Research Article

Yanqin Fan, Robert Sherman and Matthew Shum*

Estimation and Inference in an Ecological Inference Model

DOI 10.1515/jem-2015-0006
Previously published online July 9, 2015

Abstract: We interpret an ecological inference model as a treatment effects model in which the outcomes of interest and the conditional covariates come from separate datasets. In this setting, the counterfactual distributions and policy parameters of interest are only partially identified under a selection on observables assumption. In this paper, we provide estimation and inference procedures for structural prediction and counterfactual analysis in such models. We also illustrate the procedures with an application to US presidential elections.

Keywords: Copula; ecological inference; partial identification; treatment effects.

JEL Codes: C14; C31.

1 Introduction

Combining aggregate and individual-level data arises naturally in many empirical applications. These include the "classic" *ecological inference* (EI) problem from political science on predicting (individual) voting behavior in which one dataset consists of aggregate vote shares and the other consists of individual's characteristics; demand analysis in economics in which one dataset consists of market shares and the other consists of consumer demographics; poverty analysis in which one dataset consists of program participation and the other consists of demographic attributes.

Early work on partial identification in the EI problem focuses on predicting individual-level behavior when both outcomes and individual characteristics are binary. See Duncan and Davis (1953) and Goodman (1953). King (1997) and Cross and Manski (1999, 2002) formalize and extend this work to the case of *continuous* outcome variables with *discrete* individual characteristics. Recently, Fan, Sherman, and Shum (2014) generalize this model to the case where either outcomes or covariates can be continuous or discrete. In order to do this, they develop a novel approach, which borrows and combines tools and insights from the treatment effect and copula literatures. Specifically, they reinterpret the forecasting problem in Cross and Manski (1999) as a program evaluation problem. By doing so, the structural prediction exercise in the EI model becomes analogous to a problem of identifying mean counterfactual outcomes in a treatment effect model. However, the outcomes of interest are observed only at the aggregate level, while the conditioning covariates are observed at the individual-level. Because the data are contained in two separate datasets, the usual formulas for the mean counterfactual outcomes cannot be used. However, by applying the classical monotone rearrangement inequality, they obtain *bounds* on these counterfactual outcomes. Other

***Corresponding author: Matthew Shum,** Caltech, HSS, 1200 East California Blvd., Pasadena, CA 91125, USA, E-mail: mshum@caltech.edu
Yanqin Fan: Department of Economics, University of Washington, Box 353330, Seattle, WA 98195, USA
Robert Sherman: Caltech, HSS, 1200 East California Blvd., Pasadena, CA 91125, USA

parameters, such as counterfactual distributions and average treatment effect parameters, can be similarly bounded.

The lower and upper bounds of mean counterfactual outcomes and average treatment effect parameters established in Fan, Sherman, and Shum (2014) are functionals of quantile functions of generated variables. In this paper we provide plug-in estimators of these bounds and establish asymptotic properties of the plug-in estimators by making heavy use of modern empirical process methods similar to Linton, Song, and Whang (2010). A numerical example and a simple application to recent US presidential elections illustrate the methods.

The rest of this paper is organized as follows. Section 2 introduces the EI model, the traditional approach to the structural prediction or counterfactual analysis in the EI model and its limitations, and our new approach. In Section 3, we review the main identification results of Fan, Sherman, and Shum (2014). Section 4 presents estimators of the bounds on mean counterfactual outcomes and results from a numerical example. An empirical application is presented in Section 5. Section 6 establishes asymptotic properties of the plug-in estimators of the bounds on mean counterfactual outcomes. The last section concludes. Technical proofs are presented in the Appendix.

Throughout the rest of this paper, we use $F_{A|B}(\cdot|b)$ and $f_{A|B}(\cdot|b)$ to denote the distribution function and density function of the random variable A conditional on $B=b$. For a distribution function F, we use $F^{-1}(\cdot)$ to denote its quantile function.

2 Ecological Inference Model

2.1 Set-up and the Traditional EI Approach

We start by describing the formal setup we use for the EI model. Since our paper departs from the standard EI tradition (e.g. King [1997] and Cross and Manski [1999, 2002]) and applies tools and insights from the treatment effect literature to the EI problem, we will present three running examples.

Example A: vote outcomes across election years. This is the "classic" ecological inference problem from political science. Let D be the indicator for election year, Y_D be an indicator for the party voted for in election year D, and Z denote a vector of demographic variables for voters in the different election years. In this setting, the counterfactual outcome distributions are of particular interest. For example, for US presidential election data, the counterfactual distribution $F_{Y_{2000}|D}(\cdot|1980)$ describes how the 1980 US population would have voted if they had to choose among the candidates (GW Bush vs. Gore) in the 2000 presidential election. Indeed, Cross and Manski (1999) analyze the prediction of the counterfactual vote outcomes $E(Y_{1996}|D=2000)$, using our notation. We will also use this example in our empirical illustration below. ■

Example B: changes in wage distribution across time. This example is drawn from DiNardo, Fortin, and Lemieux (1996). As in the voting example above, D is a binary indicator for two different years: $D=0$ for the baseline year 1988, and $D=1$ for the counterfactual year 1979. Y_D denotes wages in year D, and DiNardo, Fortin and Lemieux focus on estimating $f_{Y_0|D}(\cdot|1)$, which they interpret as the counterfactual density of wages "if individual attributes had remained at their 1979 levels and workers had been paid according to the wage schedule observed in 1988."[1] ■

1 DiNardo, Fortin, and Lemieux (1996) observe the variables (Y, D, Z) in the same data set. We reference their work because they interpret time as a treatment variable in their counterfactual analyses; we interpret time in the same way in our empirical illustration in Section 5.

Example C: demand across different geographic markets. Let D be an indicator for geographic market, and $Y_D \in \{0, 1, ..., J\}$ be a multinomial indicator of choice among one of the J competing products (with $Y_D=0$ indicating purchase of no product). Z denotes a vector of demographic variables for consumers in the different geographic markets. In this setting, the prices of the products P_D vary over geographic markets, so that the counterfactual distribution $F_{Y_d|D}(\cdot|d')$ can be interpreted as the counterfactual demand of the consumers in market d (who faced prices P_d) if their prices were changed to P_d', the prices prevailing in market d'. ■

To incorporate all three examples in our set-up, we let $D \in \{0, 1\}$ denote two time periods or geographic markets ("treatments"), and let Y_D denote the corresponding outcome variable of interest for $D=0, 1$. Using the treatment effect interpretation, Y_0 and Y_1 are considered "potential outcomes." That is, each individual agent, regardless of the treatment group that he belongs to, is associated with values for *both* of the potential outcomes Y_0 and Y_1. However, because each agent belongs to only one of the treatment groups, his observed outcome is $Y=Y_1D+Y_0(1-D)$.

In typical EI models, the outcome variables are discrete variables, taking $M \geq 2$ distinct values. In the *aggregate dataset*, we assume that the joint distribution function of (Y, D) is observed; since both of these are discrete variables, this joint distribution function takes the form of a $M \times 2$ matrix with (m,d)-element equal to the aggregate probability $\Pr(Y=m, D=d)$. The second dataset is an *individual-level* sample, containing a sample of individual-level demographic variables $Z \in \mathcal{Z} \subset R^{d_z}$ and treatment indicators D.[2]

Let $F_{Y_d|D}(\cdot|d')$ denote the conditional distribution function of Y_d given $D=d'$ and $F_{Y_d}(\cdot)$ the marginal distribution function of Y_d. In the aggregate data, the two conditional distributions $F_{Y_d|D}(\cdot|d)$ for $d=0, 1$ are observed. However, because of differences between the individuals in the two aggregate units, we want to disentangle differences between the two distributions due to inherent behavioral differences (i.e. inherent differences between Y_d and Y_d') versus demographic differences (i.e. as captured in the distribution of the covariates Z) between the two units.[3] In the EI literature (as well as the treatment effects literature), the typical parameters of interest such as mean counterfactual outcomes are (functionals of) the (i) marginal potential outcome distributions $F_{Y_d}(\cdot)$, $d=0, 1$; as well as the (ii) counterfactual outcome distributions $F_{Y_d|D}(\cdot|d')$, for $d, d' \in \{0, 1\}$, $d \neq d'$.

The **traditional EI approach** is a two-step procedure in which we first infer individual-level behavior and then draw inferences for counterfactual outcomes from the individual-level behavior, as in Cross and Manski (1999, 2002) and Cho and Manski (2008). To illustrate, suppose the individual-level covariates Z are discrete-valued and the interest is in the mean counterfactual outcome $E(Y_0|D=1)$. The EI approach starts with the identification or partial identification of the *long regression* $E(Y_0|D=0, Z=z)$ from two identified marginal distributions: $F_{Y_0|D}(\cdot|0)$ identified from the aggregate data and $F_{Z|D}(\cdot|0)$ identified from the individual data. With *the invariance assumption*: $E(Y_0|D=0, Z=z)=E(Y_0|D=1, Z=z)$, the identified set for $E(Y_0|D=1)$ can be deduced from the identified set for the long regression $E(Y_0|D=0, Z=z)$ and $F_{Z|D}(\cdot|1)$ which is identified from the individual data. This follows from

$$\begin{aligned} E(Y_0 \mid D=1) &= E[\,E(Y_0 \mid D=1, Z) \mid D=1] \\ &= E[\,E(Y_0 \mid D=0, Z) \mid D=1]. \end{aligned} \tag{1}$$

However, when the covariates Z are continuous, the identified set for the long regression $E(Y_0|D=0, Z=z)$ is no longer computable using the Cross and Manski method of stacked distributions, which requires that Z be discrete. A different approach is needed when Z is continuously distributed.

2.2 A New Approach

To allow for both discrete and continuous covariates in Z, Fan, Sherman, and Shum (2014) develop a completely new approach to the counterfactual analysis in the EI model. Their approach draws on, and combines,

2 For simplicity of notation, we assume that the only common variable to both datasets is D. It is straightforward to extend the identification results in this paper to incorporate covariates that are observable in both data sets.

3 Firpo, Fortin and Lemieux (2010) refer to these as, respectively, the "structure" and compositional effects.

tools and insights from the treatment effect literature and the copula literature. In the treatment effect literature, the *selection on observables* assumption is often invoked to identify and evaluate average treatment effects.[4] It consists of the following two conditions:

(**C1**) Let (Y_1, Y_0, D, Z) have a joint distribution. For all $z \in \mathcal{Z}$, (Y_1, Y_0) is jointly independent of D conditional on $Z=z$.

(**C2**) For all $z \in \mathcal{Z}$, $0<p(z)<1$, where $p(z)=\Pr(D=1|Z=z)$.

(C1) is a conditional independence assumption and (C2) is a support assumption. Note that (C1) strengthens the invariance assumption $E(Y_0|D=0, Z=z)=E(Y_0|D=1, Z=z)$ underlying the traditional EI approach (see above). Strengthening the invariance assumption to a full conditional independence assumption allows us to address interesting policy parameters besides the structural prediction problem in the EI framework.

Example B (cont'd): The analysis in DiNardo, Fortin, and Lemieux (1996; hereafter DFL) can be reinterpreted in light of assumption (C1); see Firpo, Fortin, and Lemieux (2010). Under this assumption, the counterfactual wage density is:

$$\begin{aligned} f_{Y_0|D}(y|1) &= \int f_{Y_0,Z|D}(y, z|1)dz \\ &= \int f_{Y_0|Z,D}(y|z, 1) f_{Z|D}(z|1)dz \\ &= \int f_{Y_0|Z,D}(y|z, 0) f_{Z|D}(z|1)dz \end{aligned} \tag{2}$$

where the third line applies (C1). Eq. (2) above corresponds directly to Eq. (3) in DFL (subject to changes in notation), showing how DFL's analysis fits into the framework described here. ■

Example C (cont'd): To understand the conditional independence assumption (C1) further, we consider a model $Y_d=g(P=P_d, Z, \eta_d)$ where η_d captures all other unobservables in demand. Note that conditional on Z, the only randomness in demand Y_d is due to η_d. The conditional independence assumption essentially boils down to $(\eta_0, \eta_1 \perp P)|Z$. This can be interpreted as an "exogeneity" assumption on prices, conditional on Z. That is, conditional on Z, the observed demographic variables, the unobservables η_d do not affect prices. This is an undoubtedly strong assumption – it rules out, for instance, unobserved demand shocks which affect firms' pricing decisions – but, as we will see, identification is already non-trivial even in this simple case. ■

When (Y, D, Z) are all observed in a single dataset (so that there is no EI problem), it is well known that, under (C1) and (C2), the conditional distribution functions of Y_d given Z denoted as $F_{Y_1|Z}(y|z)$ and $F_{Y_0|Z}(y|z)$ are identified:

$$\begin{aligned} F_{Y_1|Z}(y|z) &= P(Y_1 \le y|Z=z) = P(Y_1 \le y|Z=z, D=1) \\ &= P(Y \le y|Z=z, D=1), \end{aligned} \tag{3}$$

$$F_{Y_0|Z}(y|z) = P(Y \le y|Z=z, D=0), \tag{4}$$

and the marginal distributions $F_{Y_1}(y)$, $F_{Y_0}(y)$ are identified as well:

$$F_{Y_1}(y) = E\left[\frac{D}{p(Z)} I\{Y \le y\}\right] \text{ and } F_{Y_0}(y) = E\left[\frac{1-D}{1-p(Z)} I\{Y \le y\}\right]. \tag{5}$$

4 See, e.g. Rosenbaum and Rubin (1983a,b), Hahn (1998), Heckman et al. (1998), Dehejia and Wahba (1999), and Hirano, Imbens, and Ridder (2000), to name only a few.

Moreover under (C1) and (C2), the counterfactual distribution function $F_{Y_0|D}(y|1)$ is also identified:

$$F_{Y_0|D}(y|1)=\frac{1}{p_1}E\left[\frac{(1-D)p(Z)}{1-p(Z)}I\{Y\le y\}\right],$$

where p_d=Pr(D=d) for d=1, 0. Thus parameters that are functionals of $F_{Y_1|Z}(y|z)$, $F_{Y_0|Z}(y|z)$, $F_{Y_0|D}(y|1)$, and the marginal distributions are identified.

In the EI model, however, (Y, D) and (Z, D) are observed in separate datasets, so the conditional expectations in Eqs. (3) and (4) can no longer be identified from the available data; neither can the unconditional distributions in Eq. (5). Fan, Sherman, and Shum (2014) show that once the variables (Y, D, Z) are not simultaneously observed in the same dataset, we no longer have point identification of these distributions, even under assumptions (C1) and (C2). To tackle this problem, they resort to the Cambanis-Simons-Stout inequality (see Lemma 3.1 below) to establish sharp bounds on some of these quantities, which is reviewed in the next section.

3 Counterfactual Analysis in the EI Model

In this section, we review the main identification results of Fan, Sherman, and Shum (2014). These include identification/partial identification results for the marginal and counterfactual marginal distributions of the potential outcomes Y_0, Y_1 and for functionals of these distributions, including the traditional program evaluation parameters such as the average treatment effect and treatment effect for the treated.

Throughout the rest of this paper, we assume Assumption (I) below:

Assumption (I) Let W=1/$p(Z)$ and V=1/[1–$p(Z)$]. Assume $Var\,(W)<\infty$ and $Var(V)<\infty$. In addition, let g denote a measurable function such that $Var(g(Y_d))<\infty$ for d=1, 0.

Theorem 3.1 establishes sharp bounds on the mean of $g(Y_d)$. This result is proved in Fan, Sherman, and Shum (2014).

Theorem 3.1 (i) *Let* $\mu_d(g)\equiv E(g(Y_d))$. *Then* $\mu_d^L(g)\le\mu_d(g)\le\mu_d^U(g)$, *where* d=1, 0 *and*

$$\mu_1^L(g)=E\left[D\int_0^1 F^{-1}_{g(Y)|D}(1-u|D)F^{-1}_{W|D}(u|D)du\right],$$
$$\mu_1^U(g)=E\left[D\int_0^1 F^{-1}_{g(Y)|D}(u|D)F^{-1}_{W|D}(u|D)du\right],$$
$$\mu_0^L(g)=E\left[(1-D)\int_0^1 F^{-1}_{g(Y)|D}(1-u|D)F^{-1}_{V|D}(u|D)du\right],$$
$$\mu_0^U(g)=E\left[(1-D)\int_0^1 F^{-1}_{g(Y)|D}(u|D)F^{-1}_{V|D}(u|D)du\right].$$

Without additional information, the bounds are sharp.

(ii) *Let* $\mu_{d|1}(g)\equiv E(g(Y_d)|D=1)$. *Then* $\mu_{1|1}(g)$ *is identified:* $\mu_{1|1}(g)\equiv E(Dg(Y)/p_1)$ *and* $\mu_{0|1}^L(g)\le\mu_{0|1}(g)\le\mu_{0|1}^U(g)$, *where*

$$\mu_{0|1}^L(g)=\frac{1}{p_1}E\left[(1-D)\int_0^1 F^{-1}_{g(Y)|D}(1-u|D)F^{-1}_{\frac{V}{W}|D}(u|D)du\right],$$
$$\mu_{0|1}^U(g)=\frac{1}{p_1}E\left[(1-D)\int_0^1 F^{-1}_{g(Y)|D}(u|D)F^{-1}_{\frac{V}{W}|D}(u|D)du\right].$$

Without additional information, the bounds are sharp.

3.1 Structural Prediction and Treatment Effects

Let $\Delta=Y_1-Y_0$ denote the individual treatment effect. Let μ_Δ, $\mu_{\Delta|1}$ denote, respectively, the average treatment effect (ATE) and the average treatment effect on the treated (TT), i.e. $\mu_\Delta=E(\Delta)$ and $\mu_{\Delta|1}=E(\Delta|D=1)$. Bounds on μ_Δ and $\mu_{\Delta|1}$ follow immediately from Theorem 3.1:

$$\mu_1^L-\mu_0^U\le\mu_\Delta\le\mu_1^U-\mu_0^L, \tag{6}$$

and

$$E\left[\frac{D}{p_1}Y\right]-\mu_{0|1}^U\le\mu_{\Delta|1}\le E\left[\frac{D}{p_1}Y\right]-\mu_{0|1}^L.$$

Let $g(Y_d)=I\{Y_d\le Y\}$ in Theorem 3.1. Noting

$$F^{-1}_{I_Y|D}(u|D)=\begin{cases}0 & \text{for } u\in[0,\ 1-F_{Y|D}(y|D)]\\ 1 & \text{for } u\in[1-F_{Y|D}(y|D),\ 1]\end{cases},$$

where $I_Y=I\{Y\le y\}$, we obtain the identified sets for the marginal distribution functions of the potential outcomes, $F_{Y_1}(y)$, $F_{Y_0}(y)$ in part (i) of Theorem 3.2 below. The identified sets for the counterfactual marginal distribution functions $F_{Y_1|D}(y|1)$, and $F_{Y_0|D}(y|1)$ are obtained similarly. Theorem 3.2 is also proved in Fan, Sherman, and Shum (2014).

Theorem 3.2 (i) *For $d=0, 1$, we have: $F_d^L(y)\le F_{Y_d}(y)\le F_d^U(y)$, where*

$$\begin{aligned}
F_1^L(y)&=E[D\textstyle\int_0^{F_{Y|D}(y|D)}F^{-1}_{W|D}(u|D)du],\\
F_1^U(y)&=E[D\textstyle\int_{1-F_{Y|D}(y|D)}^{1}F^{-1}_{W|D}(u|D)du],\\
F_0^L(y)&=E[(1-D)\textstyle\int_0^{F_{Y|D}(y|D)}F^{-1}_{V|D}(u|D)du],\\
F_0^U(y)&=E[(1-D)\textstyle\int_{1-F_{Y|D}(y|D)}^{1}F^{-1}_{V|D}(u|D)du].
\end{aligned}$$

Without additional information, the bounds are sharp (both pointwise and uniformly).

(ii) *$F_{Y_1|D}(y|1)$ is identified: $F_{Y_1|D}(y|1)=E[DI\{Y\le y\}/p]$ and $F_{Y_0|D}(y|1)$ is partially identified: $F^L_{0|D}(y|1)\le F_{Y_0|D}(y|1)\le F^U_{0|D}(y|1)$, where*

$$\begin{aligned}
F^L_{0|D}(y|1)&=\frac{1}{p_1}E[(1-D)\textstyle\int_0^{F_{Y|D}(y|D)}F^{-1}_{\frac{V}{W}|D}(u|D)du],\\
F^U_{0|D}(y|1)&=\frac{1}{p_1}E[(1-D)\textstyle\int_{1-F_{Y|D}(y|D)}^{1}F^{-1}_{\frac{V}{W}|D}(u|D)du].
\end{aligned}$$

Without additional information, the bounds are sharp (both pointwise and uniformly).

The uniform sharpness of the bounds in Theorem 3.2 allows us to establish sharp bounds on monotone functionals of the marginal or counterfactual marginal distribution functions. Such functionals include the quantile treatment effects (QTE) defined as

$$\begin{aligned}
QTE_u&=F^{-1}_{Y_1}(u)-F^{-1}_{Y_0}(u)\quad\text{and}\\
QTE_{u|1}&=F^{-1}_{Y_1|D}(u|1)-F^{-1}_{Y_0|D}(u|1),
\end{aligned}$$

where $u\in(0, 1)$, and the cumulative distribution functions

$$F_\Delta(\delta)=\Pr(\Delta\le\delta),$$
$$F_\Delta(\delta|D=1)=\Pr(\Delta\le\delta|D=1),$$

and the corresponding quantile functions.

For example, it may be of interest to evaluate the probability of a positive treatment effect: either $\Pr(\Delta>0)$ or $\Pr(\Delta>0|D=1)$, and the median of Δ for the treated. Sharp bounds on these parameters can be established by extending the reach of Theorem 3.2 by applying the lemma below adapted from Frank, Nelson, and Schweizer (1987); see also Fan and Park (2009, 2010).

Lemma 3.3 *Let* $F_\Delta(\delta|\cdot)=\Pr(\Delta\le\delta|\cdot)$. *Then* $F_\Delta^L(\delta|\cdot)\le F_\Delta(\delta|\cdot)\le F_\Delta^U(\delta|\cdot)$, *where*

$$F_\Delta^L(\delta|\cdot)=\max\left(\sup_y\left[F_1(y|\cdot)-F_0(y-\delta|\cdot)\right],0\right),$$
$$F_\Delta^U(\delta|\cdot)=1+\min\left(\inf_y\left[F_1(y|\cdot)-F_0(y-\delta|\cdot)\right],0\right),$$

where $F_d(y|\cdot)=\Pr(Y_d\le y|\cdot)$ *for* $d=1, 0$.

Consider, for instance, the CDF $F_\Delta(\delta|D=1)$. From Theorem 3.2 and Lemma 3.3, we have:

$$F_\Delta^L(\delta|D=1)\le F_\Delta(\delta|D=1)\le F_\Delta^U(\delta|D=1),$$

where

$$F_\Delta^L(\delta|D=1)=\max\left(\sup_y\left[F_{Y_1|D}(y|1)-F_{0|D}^U(y-\delta|1)\right],0\right),$$
$$F_\Delta^U(\delta|D=1)=1+\min\left(\inf_y\left[F_{Y_1|D}(y|1)-F_{0|D}^L(y-\delta|1)\right],0\right).$$

Sharp bounds on the quantile function of $F_\Delta(\delta|D=1)$ follow from sharp bounds on $F_\Delta(\delta|D=1)$.

3.2 Group-Level Behavior

In some applications, group level behavior may be of interest, see Cho and Manski (2008). Let $\mathcal{S}$ denote a subset of $\mathcal{Z}$ such that $\Pr(Z\in\mathcal{S})>0$. The reach of Theorem 3.2 can be extended to establish sharp bounds on the group-level behavior for individuals with characteristics $Z\in\mathcal{S}$ even when some of the components of Z are continuous. To illustrate, consider $E(Y_0|Z\in\mathcal{S})$

$$\begin{aligned}E(Y_0|Z\in\mathcal{S})&=\frac{1}{\Pr(Z\in\mathcal{S})}E((1-D)YVI\{Z\in\mathcal{S}\})\\&=\frac{1}{\Pr(Z\in\mathcal{S})}E((1-D)E[YVI\{Z\in\mathcal{S}\}|D])\end{aligned}$$

leading to

$$\begin{aligned}&\frac{1}{\Pr(Z\in\mathcal{S})}E\left((1-D)\int_0^1F_{Y|D}^{-1}(1-u|D)F_{V_\mathcal{S}|D}^{-1}(u|D)du\right)\\&\le E(Y_0|Z\in\mathcal{S})\le\frac{1}{\Pr(Z\in\mathcal{S})}E\left((1-D)\int_0^1F_{Y|D}^{-1}(u|D)F_{V_\mathcal{S}|D}^{-1}(u|D)du\right),\end{aligned}$$

where $V_\mathcal{S}=VI\{Z\in\mathcal{S}\}$.

4 Estimation of the Bounds

In describing the estimators, we simplify further to the case where the outcomes are binary: Y_0, $Y_1 \in \{0, 1\}$.[5] Let $P_{01}=\Pr(Y=0|D=1)$ and $P_{00}=\Pr(Y=0|D=0)$. Thereby, the first (aggregate) dataset contains two values P_{01} and P_{00}, while the second (individual-level) dataset contains $(Z_i, D_i)_{i=1}^n$, where i denotes individual. The asymptotic properties of all the estimators are derived under the assumption that P_{01} and P_{00} are completely known (observed without sampling error). It is a straightforward, but tedious, extension to consider the case when P_{01} and P_{00} are observed with noise.

In this section we present our estimators of the bounds on mean counterfactual outcomes (corresponding to the bounds defined in Theorem 3.1 above), and a numerical example. We delay the detailed discussion of the asymptotic theory for these estimators to Section 6 below. Moreover, estimation of other functionals of the counterfactual outcome distribution (as in Theorem 3.2) proceeds in analogous fashion as for the mean; for this reason, we do not discuss this in the paper, but details are available from the authors upon request.

4.1 Estimators

We begin by considering the estimation of the (bounds on) mean counterfactual outcomes. For binary outcomes,

$$F_{Y|D}^{-1}(1-u|D)=\begin{cases}0 & \text{if } 1-\Pr(Y=0|D)\le v\le 1\\ 1 & \text{if } 0\le u<1-\Pr(Y=0|D)\end{cases}$$

so an application of Theorem 3.1 yields:

$$\mu_1^L=E[D\int_0^{1-\Pr(Y=0|D)}F_{W|D}^{-1}(u|D)du]=p_1\int_0^{1-P_{01}}F_{W|D}^{-1}(u|1)du,$$
$$\mu_1^U=E[D\int_{\Pr(Y=0|D)}^1F_{W|D}^{-1}(u|D)du]=p_1\int_{P_{01}}^1F_{W|D}^{-1}(u|1)du,$$

and

$$\mu_0^L=E[(1-D)\int_0^{1-\Pr(Y=0|D)}F_{V|D}^{-1}(u|D)du]=(1-p_1)\int_0^{1-P_{00}}F_{V|D}^{-1}(u|0)du,$$
$$\mu_0^U=E[(1-D)\int_{\Pr(Y=0|D)}^1F_{V|D}^{-1}(u|D)du]=(1-p_1)\int_{P_{00}}^1F_{V|D}^{-1}(u|0)du.$$

Similarly, we have:

$$\mu_{0|1}^L=\frac{1-p_1}{p_1}\int_0^{1-P_{00}}F_{V/W|D}^{-1}(u|0)du,$$
$$\mu_{0|1}^U=\frac{1-p_1}{p_1}\int_{P_{00}}^1F_{V/W|D}^{-1}(u|0)du.$$

To estimate these bounds, we need to estimate p_1 and the quantile functions: $F_{W|D}^{-1}(u|d)$, $F_{\frac{V}{W}|D}^{-1}(u|d)$, and $F_{V|D}^{-1}(u|d)$ for d=1, 0. An estimator of p_1 is $\hat{p}_1=n^{-1}\sum_{i=1}^n D_i$. To estimate the quantile functions, we let

$$F_{P|D}(p|d)=\Pr(p(Z)\le p|D=d)$$

and, assuming $p(Z)$ is a continuous random variable, we obtain:

5 The extension to more values of the outcomes, or to more than two treatments, is straightforward.

$$F_{W|D}(w|d)=\Pr\left(p(Z)\geq\frac{1}{w}|D=d\right)=1-F_{P|D}\left(\frac{1}{w}|d\right),$$
$$F_{V|D}(v|d)=\Pr\left(p(Z)\leq 1-\frac{1}{v}|D=d\right)=F_{P|D}\left(1-\frac{1}{v}|d\right),$$

and

$$F_{V/W|D}(a|d)=\Pr\left(p(Z)\leq\frac{a}{1+a}|D=d\right)=F_{P|D}\left(\frac{a}{1+a}|d\right).$$

Let $Q_{A|D}(u|d)\equiv F^{-1}_{A|D}(u|d)$ and $\hat{Q}_{A|D}(u|d)$ denote a consistent estimator of $Q_{A|D}(u|d)$ for $A=W$, V/W, V. Then

$$Q_{W|D}(u|d)=\frac{1}{F^{-1}_{P|D}(1-u|d)},\quad Q_{V|D}(u|d)=\frac{1}{1-F^{-1}_{P|D}(u|d)},\quad Q_{V/W|D}(u|d)=\frac{F^{-1}_{P|D}(u|d)}{1-F^{-1}_{P|D}(u|d)}.$$

Let $\hat{p}(z)$ denote any consistent estimator of the propensity score $p(z)$ using dataset $\{Z_i, D_i\}_{i=1}^n$. Further let

$$\hat{P}_i=\hat{p}(Z_i),\ i=1,\ \ldots,\ n.$$

Our quantile estimators are constructed from the generated dataset: $\{\hat{P}_i, D_i\}_{i=1}^n$. Let

$$\hat{F}^{-1}_{P|D}(u|d)=\inf\{a\in\mathcal{A}:\hat{F}_{P|D}(a|d)\geq u\},\tag{7}$$

Where $\hat{F}_{P|D}(a|d)$ is defined as:

$$\hat{F}_{P|D}(a|d)=\frac{n^{-1}\sum_{i=1}^n I\{\hat{P}_i\leq a\}I\{D_i=d\}}{\hat{p}_d}.\tag{8}$$

Using the quantile estimator $\hat{F}^{-1}_{P|D}(u|d)$, we construct the following quantile estimators:

$$\hat{Q}_{W|D}(u|d)=\frac{1}{\hat{F}^{-1}_{P|D}(1-u|d)},\quad \hat{Q}_{V|D}(u|d)=\frac{1}{1-\hat{F}^{-1}_{P|D}(u|d)},\quad \hat{Q}_{V/W|D}(u|d)=\frac{\hat{F}^{-1}_{P|D}(u|d)}{1-\hat{F}^{-1}_{P|D}(u|d)},$$

and propose the estimators of the bounds below:

$$\hat{\mu}^L_1=\hat{p}_1[\int_0^{1-P_{01}}\hat{Q}_{W|D}(u|1)du],$$
$$\hat{\mu}^U_1=\hat{p}_1[\int_{P_{01}}^1\hat{Q}_{W|D}(u|1)du],$$
$$\hat{\mu}^L_0=(1-\hat{p}_1)[\int_0^{1-P_{00}}\hat{Q}_{V|D}(u|0)du],$$
$$\hat{\mu}^U_0=(1-\hat{p}_1)[\int_{P_{00}}^1\hat{Q}_{V|D}(u|0)du],$$

and

$$\hat{\mu}^L_{0|1}=\frac{1-\hat{p}_1}{\hat{p}_1}\int_0^{1-P_{00}}\hat{Q}_{V/W|D}(u|0)du,$$
$$\hat{\mu}^U_{0|1}=\frac{1-\hat{p}_1}{\hat{p}_1}\int_{P_{00}}^1\hat{Q}_{V/W|D}(u|0)du.$$

4.2 Numerical Comparisons from a Test Model

For assessing the performance of our bounds, we consider a simple simulation example. We consider the following test model:

$$\begin{aligned} &Y_1^*=\gamma_1 Z+V_1, \quad Y_0^*=\gamma_0 Z+V_0, \\ &Y_1=\frac{\exp(Y_1^*)}{1+\exp(Y_1^*)}, \quad Y_0=\frac{\exp(Y_0^*)}{1+\exp(Y_0^*)}, \end{aligned} \tag{9}$$

$$D=I\{Z\delta-\epsilon\geq 0\}, \tag{10}$$

in which $(Z, V_1, V_0, \epsilon)\sim N(0, I_4)$.

For this model, we can compute our bounds for various values of the parameters $(\delta, \gamma_1, \gamma_0)$. For the average treatment effects, we compare our bounds to Manski's (1990) "worst case" benchmark, which are for the case when only the aggregate data are available. Specifically, suppose $Y_1\in[0, 1]$. Then Manski's worst case bounds on μ_1 are:

$$E[Y|D=1]\,p_1\leq\mu_1\leq E[Y|D=1]\,p_1+[1-p_1].$$

The bounds are presented graphically in Figures 1 and 2. In these graphs, the solid blue line in the middle corresponds to the actual values of the ATE or ATT, which can be directly computed from the simulated data. One noteworthy feature in the ATE graphs in Figure 1 is how much our bounds shrink relative to the Manski bounds. When $\delta=0$, our bounds point identify ATE and ATT.

Subsequently, we also consider the small-sample properties of our bounds. In Table 1, we show results from a Monte Carlo study. As expected, increasing the sample size leads to more precise estimates of the bounds (but, of course, cannot tighten the bounds).

5 An Empirical Example

In this section, we consider a simple empirical example to illustrate our approach. We compute bounds on counterfactual voteshares for the US presidential elections of 1980 and 2000. We define a binary outcome variable Y_i for individual eligible voter i:

$$Y_i=\begin{cases}1 & \text{if } i \text{ votes Republican}\\ 0 & \text{if } i \text{ does not vote Republican.}\end{cases}$$

Note that $Y_i=0$ encompasses voting Democratic, voting for a third party candidate, or abstaining. Likewise, the variable D_i is a binary indicator defined as being $=1$ $(=0)$ if person i was randomly sampled from the eligible voting population in 2000 (in 1980).[6] Accordingly, we define the two potential outcome variables: for $D=0, 1$,

$$Y_{D,i}=\begin{cases}1 & \text{if } i \text{ votes Republican in election year } D\\ 0 & \text{if } i \text{ does not vote Republican in election year } D\end{cases}$$

and the relationship between Y_i and the potential outcomes is:

$$Y_i=\mathbb{1}_{D_i=1}Y_{1,i}+(1-\mathbb{1}_{D_i=0})Y_{0,i}.$$

6 Note that it is possible (though highly unlikely) for the same person to be sampled in both 2000 and 1980. However, D_i can never be equal to both zero and one since our methods formally treat each person sampled in 2000 as different from each person sampled in 1980.

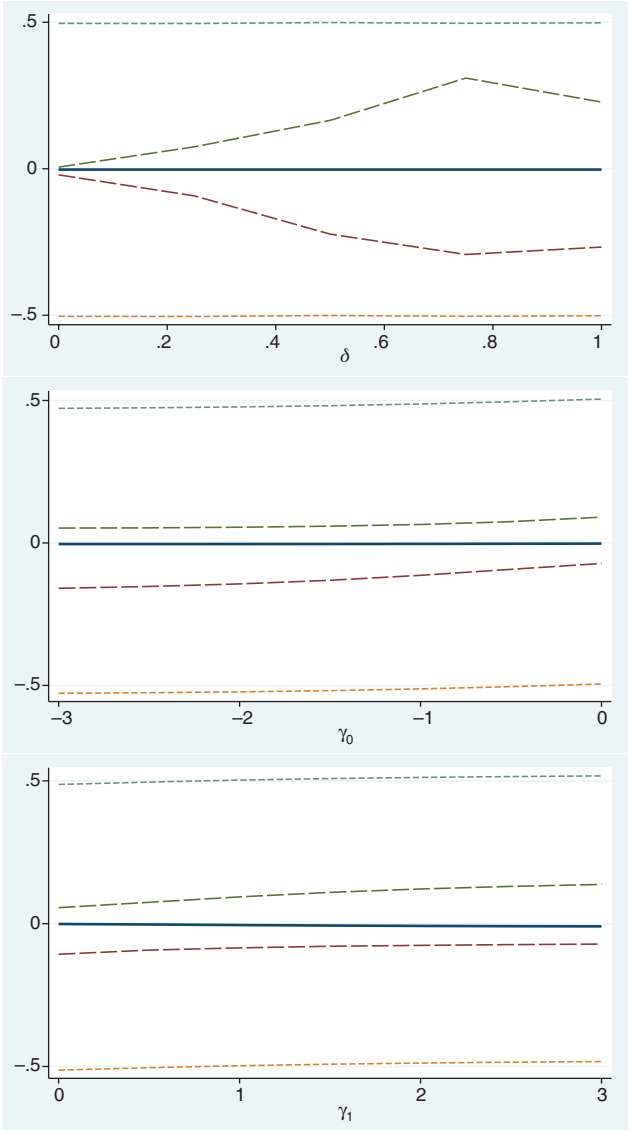

Figure 1: Test Model: Effects of Parameters on ATE Bounds.
Solid line, true ATE; Dashed lines, our bounds; Short-dashed lines, Manski (1990) bounds.
Baseline parameter values: δ=0.25; γ_0=−0.5; γ_1=0.5.

The propensity score was estimated using a probit regression on a random sample of individual-level census data from the 1980 and 2000 censuses.[7] The specification was:[8]

$$P(D_i=1|Z_i)=\Phi(\beta_0+\beta_1 LOGINC_i+\beta_2 COLLEGE_i+\beta_3 AGE_i \\ +\beta_4 NONWHITE_i+\beta_5 UNEMP_i+\beta_6 EMPL_i)$$

where $\Phi(\cdot)$ denotes the standard normal CDF. The definitions and summary statistics for the demographic variables are given in Table 2. This propensity score regression was estimated separately for each state, as well as for the nation as a whole.

7 See, e.g. Wooldridge (2010, chapter 21, section 3.3, pp. 920–923) for a discussion of methods and guidelines for estimating the propensity score parametrically, as well as references to articles indicating how to estimate the propensity score both parametrically and nonparametrically.

8 The results were quite stable to different definitions and specifications of the conditioning variables in the propensity scores.

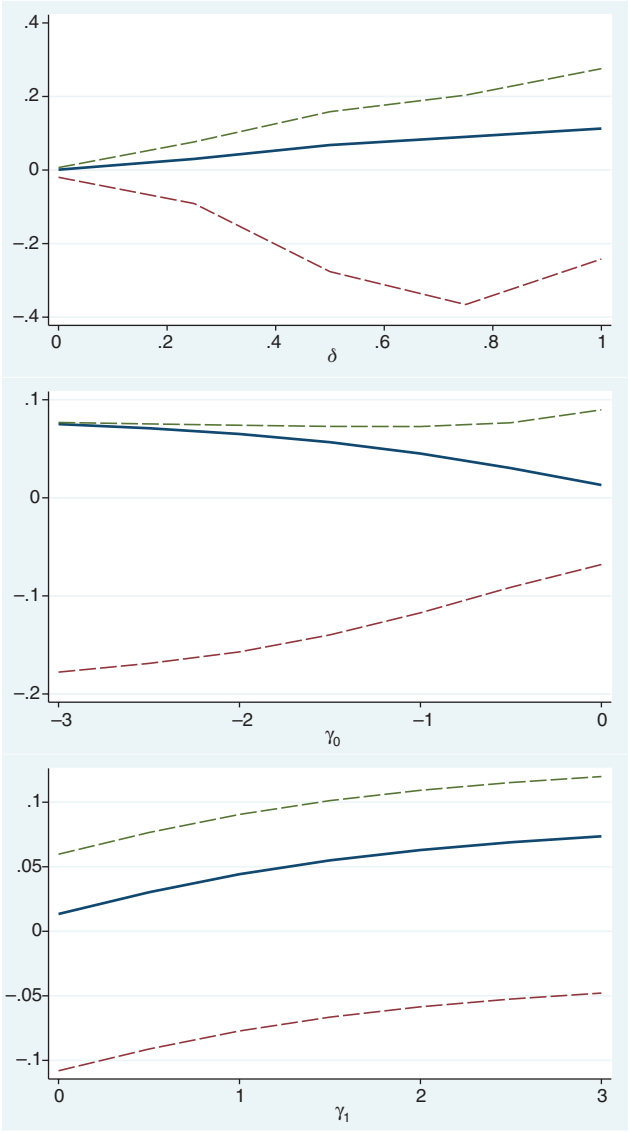

Figure 2: Test Model: Effects of Parameters on ATT Bounds.
Solid line, true ATT; Dashed lines, our bounds.
Baseline parameter values: δ=0.25; γ_0=−0.5; γ_1=0.5.

To aid in the interpretation of the results, we make a simplifying expository assumption that the main intrinsic difference across election years is the identity of the presidential candidates – Ronald Reagan and Jimmy Carter in 1980, and George W. Bush (GWB) and Al Gore in 2000. In this setting, the counterfactual outcomes $Y_{1,i}$ (resp. $Y_{0,i}$) indicate whether voter i would have voted for GWB in the 2000 election (resp. Reagan in the 1980 election). Moreover, the counterfactual probability $P(Y_0=1|D=1)$ provides an answer to the question: if Reagan and Carter had run in 2000, would an average voter in 2000 have voted for the Republican ticket? Analogously, the average treatment effect $E(Y_1-Y_0|D=1)$ answers the question: fixing the voting population in the year 2000, how would the vote share have changed if Reagan and Carter had run in 2000, instead of GWB and Gore?

5.1 Estimates of the Bounds

Bounds on these treatment effects are presented in Table 3. We calculated the bounds for each state, as well as for the nation as a whole. Looking at the TT column in the table, we see that, for all the states, the

Table 1: Test Model: Small-Sample Performance.

		Fixed parameters: δ=0.25; γ_0=−0.5; γ_1=0.5
	# obs	**[Low-bound, up-bound]**
Average TE:		
True		−0.0215
Our bounds	500	[−0.0852, 0.0849]
		[(0.0298), (0.0277)]
	1000	[−0.0855, 0.0857]
		[(0.0197), (0.0197)]
	2000	[−0.0844, 0.0848]
		[(0.0133), (0.0136)]
Manski bounds	500	[−0.5000, 0.5000]
		[(0.0100), (0.0100)]
	1000	[−0.4999, 0.5001]
		[(0.0072), (0.0072)]
	2000	[−0.4999, 0.5001]
		[(0.0050), (0.0050)]
Average TT:		
True		0.0135
Our bounds	500	[−0.0836, 0.0867]
		[(0.0290), (0.0272)]
	1000	[−0.0831, 0.0879]
		[(0.0201), (0.0200)]
	2000	[−0.0820, 0.0873]
		[(0.0133), (0.0140)]

True values computed by simulation. Propensity scores estimated using linear probability model. Reported bounds are averages over 1000 replications. Standard deviations (across replications) in parentheses.

Table 2: Summary Statistics: Variables Included in Propensity Score.

Variables		**1980**		**2000**	
		Mean	**Std. dev.**	**Mean**	**Std. dev**
Loginc	Log household per-capita annual income	9.3457	0.8223	9.7317	0.8973
Age		43.1512	17.9048	46.1245	17.6180
College	=1 if at least some college, 0 otherwise	0.3185	0.4659	0.4252	0.4944
Nonwhite	=1 if nonwhite	0.1280	0.3343	0.2030	0.4022
Unempl	=1 if unemployed	0.0388	0.1932	0.0311	0.1735
Empl[a]	=1 if employed	0.6061	0.4886	0.6268	0.4837
#obs		50,000		50,000	

[a]Omitted employment category: out of the labor force.

bounds on ATT=$E(Y_1-Y_0|D=1)$ contain zero. Moreover, for all the states, and for the nation as a whole, the actual change in Republican voteshare between election years (in the "Diff." column) lies between the upper and lower bounds of the ATT. This implies that, using the previous terminology: if we were to hold fixed the candidates in the 2000 election, we cannot reject the hypothesis that there would have been no discernible change in the vote outcomes.

To get more insight, we decompose this treatment effect into

$$\text{ATT}=P(Y_1=1|D=1)-P(Y_0=1|D=1)$$

The first term on the right-hand side is directly observed in the data, and reported in the column labelled "2000" in Table 3. We have derived bounds for the second term on the right-hand side, which are reported in Table 4 (the lower and upper bounds are, respectively, in the "$\mu^L_{0|1}$" and "$\mu^U_{0|1}$" columns). For all states,

Table 3: Changes in Republican Voteshares in US Presidential Elections, 1980/2000.

State[a]		Actual Repub. Voteshares			ATE		ATT	
		1980	2000	Diff.	Lower bound	Upper bound	Lower bound	Upper bound
1	CT	0.2937	0.2179	−0.0757	−0.1936	0.0554	−0.1441	0.1239
2	ME	0.2944	0.2927	−0.0017	−0.1547	0.1597	−0.0851	0.2124
3	MA	0.2470	0.1802	−0.0668	−0.1713	0.0490	−0.1289	0.1102
4	NH	0.3301	0.2925	−0.0376	−0.1760	0.1234	−0.1102	0.1753
5	RI	0.2181	0.1622	−0.0559	−0.1574	0.0813	−0.1142	0.1158
6	VT	0.2563	0.2582	0.0019	−0.1707	0.1725	−0.1219	0.2279
7	DE	0.2582	0.2312	−0.0270	−0.1573	0.0850	−0.1226	0.1577
8	NJ	0.2853	0.2020	−0.0834	−0.1974	0.0458	−0.1548	0.1083
9	NY	0.2238	0.1675	−0.0562	−0.1363	0.0310	−0.0985	0.0822
10	PA	0.2578	0.2433	−0.0145	−0.1234	0.1007	−0.0709	0.1464
11	IL	0.2870	0.2192	−0.0678	−0.1624	0.0361	−0.1151	0.0894
12	IN	0.3232	0.2752	−0.0480	−0.1544	0.0604	−0.1100	0.1163
13	MI	0.2940	0.2647	−0.0293	−0.1343	0.0854	−0.0814	0.1332
14	OH	0.2854	0.2768	−0.0086	−0.1066	0.0899	−0.0535	0.1397
15	WI	0.3230	0.3080	−0.0150	−0.1427	0.1137	−0.0820	0.1769
16	IA	0.3229	0.2887	−0.0343	−0.1731	0.1045	−0.1266	0.1657
17	KS	0.3283	0.3138	−0.0145	−0.1315	0.0986	−0.0764	0.1556
18	MN	0.2981	0.3031	0.0051	−0.1242	0.1376	−0.0568	0.1881
19	MO	0.3008	0.2839	−0.0168	−0.1165	0.0888	−0.0627	0.1380
20	NE	0.3717	0.3428	−0.0289	−0.1727	0.1117	−0.1247	0.1854
21	ND	0.4161	0.3636	−0.0526	−0.1851	0.1213	−0.0817	0.1856
22	SD	0.4074	0.3437	−0.0637	−0.1701	0.0447	−0.1146	0.0970
23	VA	0.2520	0.2668	0.0148	−0.0992	0.1280	−0.0408	0.1733
24	AL	0.2376	0.2822	0.0446	−0.0684	0.1568	−0.0023	0.1954
25	AR	0.2481	0.2362	−0.0119	−0.0880	0.0631	−0.0367	0.1104
26	FL	0.2702	0.2336	−0.0366	−0.1376	0.0704	−0.0856	0.1163
27	GA	0.1690	0.2329	0.0638	−0.0238	0.1522	0.0447	0.1837
28	LA	0.2724	0.2851	0.0127	−0.0956	0.1249	−0.0371	0.1634
29	MS	0.2565	0.2758	0.0193	−0.0832	0.1281	−0.0213	0.1644
30	NC	0.2142	0.2656	0.0514	−0.0544	0.1528	0.0075	0.1918
31	SC	0.1997	0.2599	0.0602	−0.0320	0.1530	0.0334	0.1907
32	TX	0.2481	0.2510	0.0029	−0.0948	0.1010	−0.0386	0.1414
33	KY	0.2451	0.2850	0.0400	−0.0886	0.1665	−0.0234	0.2038
34	MD	0.2209	0.2049	−0.0160	−0.1373	0.1079	−0.0973	0.1541
35	OK	0.3160	0.2899	−0.0261	−0.1112	0.0628	−0.0590	0.1103
36	TN	0.2374	0.2459	0.0085	−0.0989	0.1194	−0.0407	0.1615
37	WV	0.2393	0.2395	0.0001	−0.0899	0.1029	−0.0364	0.1421
38	AZ	0.2690	0.2044	−0.0647	−0.1702	0.0511	−0.1225	0.1010
39	CO	0.3079	0.2720	−0.0359	−0.1571	0.1032	−0.1007	0.1528
40	ID	0.4514	0.3601	−0.0913	−0.2133	0.0578	−0.1362	0.1174
41	MT	0.3699	0.3556	−0.0142	−0.1067	0.0863	−0.0345	0.1293
42	NV	0.2573	0.1984	−0.0589	−0.1462	0.0550	−0.0841	0.0990
43	NM	0.2793	0.2174	−0.0619	−0.1744	0.0608	−0.1304	0.1080
44	UT	0.4704	0.3354	−0.1350	−0.2414	0.0174	−0.1510	0.0829
45	WY	0.3347	0.4043	0.0696	−0.1012	0.2247	−0.0312	0.2838
46	CA	0.2578	0.1837	−0.0741	−0.1886	0.0441	−0.1536	0.0923
47	OR	0.2965	0.2749	−0.0216	−0.1407	0.1075	−0.0699	0.1607
48	WA	0.2849	0.2509	−0.0340	−0.1540	0.1170	−0.0848	0.1607
49	AK	0.3106	0.3808	0.0702	−0.0810	0.2196	0.0002	0.2560
50	HI	0.1870	0.1498	−0.0372	−0.0977	0.0309	−0.0549	0.0811
52	Natl	0.2671	0.2396	−0.0275	−0.1306	0.0812	−0.0785	0.1280

[a]State numbers correspond to y-axis labels in Figures 3 and 4.

Table 4: Example: Bounds for Marginal and Counterfactual Outcome Distributions.

State		Bounds: marginal mean outcome				Bounds: counterfct mean	
		μ_0^L	μ_0^U	μ_1^L	μ_1^U	$\mu_{0\|1}^L$	$\mu_{0\|1}^U$
1	CT	0.1713	0.3010	0.1074	0.2267	0.0940	0.3621
2	ME	0.1678	0.3066	0.1520	0.3276	0.0803	0.3778
3	MA	0.1375	0.2496	0.0783	0.1865	0.0700	0.3091
4	NH	0.1889	0.3488	0.1728	0.3123	0.1172	0.4027
5	RI	0.1146	0.2157	0.0583	0.1959	0.0464	0.2764
6	VT	0.1136	0.2979	0.1272	0.2861	0.0303	0.3801
7	DE	0.1373	0.2849	0.1275	0.2223	0.0735	0.3538
8	NJ	0.1686	0.2925	0.0951	0.2144	0.0936	0.3567
9	NY	0.1328	0.2165	0.0802	0.1638	0.0853	0.2661
10	PA	0.1569	0.2566	0.1332	0.2576	0.0969	0.3142
11	IL	0.1857	0.2835	0.1211	0.2218	0.1298	0.3343
12	IN	0.2196	0.3257	0.1713	0.2800	0.1589	0.3852
13	MI	0.1910	0.2921	0.1578	0.2764	0.1315	0.3462
14	OH	0.1894	0.2797	0.1732	0.2794	0.1371	0.3303
15	WI	0.2034	0.3303	0.1876	0.3172	0.1310	0.3899
16	IA	0.2063	0.3362	0.1631	0.3108	0.1230	0.4153
17	KS	0.2226	0.3301	0.1986	0.3213	0.1582	0.3901
18	MN	0.1815	0.3040	0.1798	0.3192	0.1150	0.3599
19	MO	0.2032	0.2943	0.1778	0.2920	0.1459	0.3466
20	NE	0.2452	0.3896	0.2169	0.3569	0.1575	0.4675
21	ND	0.2799	0.4024	0.2173	0.4012	0.1779	0.4453
22	SD	0.3055	0.4046	0.2346	0.3502	0.2467	0.4583
23	VA	0.1445	0.2580	0.1588	0.2725	0.0935	0.3077
24	AL	0.1384	0.2349	0.1666	0.2952	0.0868	0.2845
25	AR	0.1623	0.2351	0.1472	0.2254	0.1258	0.2729
26	FL	0.1637	0.2745	0.1370	0.2341	0.1172	0.3192
27	GA	0.0807	0.1570	0.1332	0.2328	0.0492	0.1882
28	LA	0.1747	0.2697	0.1742	0.2996	0.1217	0.3222
29	MS	0.1616	0.2493	0.1661	0.2897	0.1114	0.2971
30	NC	0.1172	0.2130	0.1586	0.2700	0.0738	0.2581
31	SC	0.1071	0.1904	0.1584	0.2601	0.0692	0.2265
32	TX	0.1504	0.2474	0.1526	0.2514	0.1097	0.2896
33	KY	0.1399	0.2500	0.1614	0.3064	0.0812	0.3084
34	MD	0.1115	0.2358	0.0985	0.2194	0.0508	0.3022
35	OK	0.2271	0.3033	0.1921	0.2899	0.1796	0.3489
36	TN	0.1348	0.2381	0.1392	0.2543	0.0844	0.2866
37	WV	0.1477	0.2284	0.1386	0.2507	0.0973	0.2759
38	AZ	0.1498	0.2836	0.1134	0.2009	0.1034	0.3269
39	CO	0.1861	0.3174	0.1603	0.2893	0.1192	0.3727
40	ID	0.3220	0.4488	0.2355	0.3798	0.2426	0.4962
41	MT	0.2761	0.3527	0.2460	0.3625	0.2263	0.3901
42	NV	0.1336	0.2586	0.1125	0.1886	0.0994	0.2825
43	NM	0.1661	0.2917	0.1174	0.2269	0.1094	0.3478
44	UT	0.3273	0.4569	0.2156	0.3448	0.2525	0.4864
45	WY	0.2042	0.3586	0.2574	0.4289	0.1205	0.4354
46	CA	0.1459	0.2767	0.0881	0.1901	0.0914	0.3373
47	OR	0.1778	0.2976	0.1569	0.2853	0.1142	0.3448
48	WA	0.1572	0.2890	0.1351	0.2742	0.0902	0.3358
49	AK	0.1861	0.3264	0.2454	0.4057	0.1247	0.3806
50	HI	0.1010	0.1726	0.0749	0.1319	0.0687	0.2047
52	Natl	0.1643	0.2676	0.1370	0.2455	0.1115	0.3181

these bounds contain the actual Republican voteshare in 1980 ($E(Y_0)$, as reported in Table 3). These results show that we cannot reject the hypothesis that the candidates did not make a difference, in the sense that the Republican voteshares would have been the same in 2000 even if Reagan and Carter were running instead of Bush and Gore.

While these bounds are inconclusive, they shed light on one hypothesis, which is that appeal for Reagan, and his anti-government message, was especially strong in 1980 when the economy was in a recession. Our results show that the data do not allow us to reject the hypothesis that Reagan's message would have resonated just as well with the voters in 2000, a year of economic prosperity (as evidenced in the summary statistics in Table 2).

5.2 Confidence Sets

All the parameters of interest including the mean outcomes μ_1, μ_0, the counterfactual outcome $\mu_{0|1}$, and the treatment effect parameters μ_Δ, $\mu_{\Delta|1}$ are interval identified and estimators of their bounds are jointly asymptotically normally distributed. Their inference falls within the framework of Andrews and Soares (2010), Fan and Park (2010), and Stoye (2009), and Chernozhukov, Hong, and Tamer (2007).

To compute the confidence sets for ATE and ATT, we followed the procedure in Stoye (2009). As an input into Stoye's procedure, we computed the standard errors and correlation between the lower and upper bound estimates for each of the five parameters $E(Y_0)$, $E(Y_1)$, $E(Y_0|D=1)$, $ATE=E(Y_1)-E(Y_0)$, and $ATT=E(Y_1|D=1)-E(Y_0|D=1)$. Because the asymptotic variances of our estimators of the bounds on these parameters are very complicated due to the use of generated variables, we used bootstrap simulation to approximate the variances and correlations.[9]

In Figures 3 and 4, we present graphs of the estimated bounds, along with the confidence sets. We see that, across all the states, the confidence sets and the estimated bounds are very close, suggesting that sampling error is not a big concerns in these bounds estimates. Overall, then, this suggests that sampling error due to estimation is minor relative to the magnitude of the differences between the upper and lower bounds for the parameters.

6 Asymptotics for Bounds Estimators

In this section we present the asymptotic theory for the bounds estimators of the mean counterfactual outcomes defined in Section 4 above. Let $\mathcal{W}$, $\mathcal{V}$, and $\mathcal{A}$ denote the support of the random variables W, V, and V/W respectively. To simplify the derivation of the asymptotic properties of our estimators, we introduce the following distribution functions:

$$G_{W|D}(a|d)=F_{W|D}\left(\frac{a}{p_1}|d\right),\ \frac{a}{p_1}\in\mathcal{W}$$
$$G_{V|D}(a|d)=F_{V|D}\left(\frac{a}{1-p_1}|d\right),\ \frac{a}{1-p_1}\in\mathcal{V}$$
$$G_{V/W|D}(a|d)=F_{V/W|D}\left(\frac{ap_1}{1-p_1}|d\right),\ \frac{ap_1}{1-p_1}\in\mathcal{A}$$

and their estimators:

9 When the P's are estimated from a separate sample independent of $(Z_i, D_i)_{i=1}^n$, the bootstrap procedure will include drawing bootstrap samples from each sample and re-estimating the P's and the conditional quantiles involved using the two bootstrap samples. We used 100 bootstrap samples.

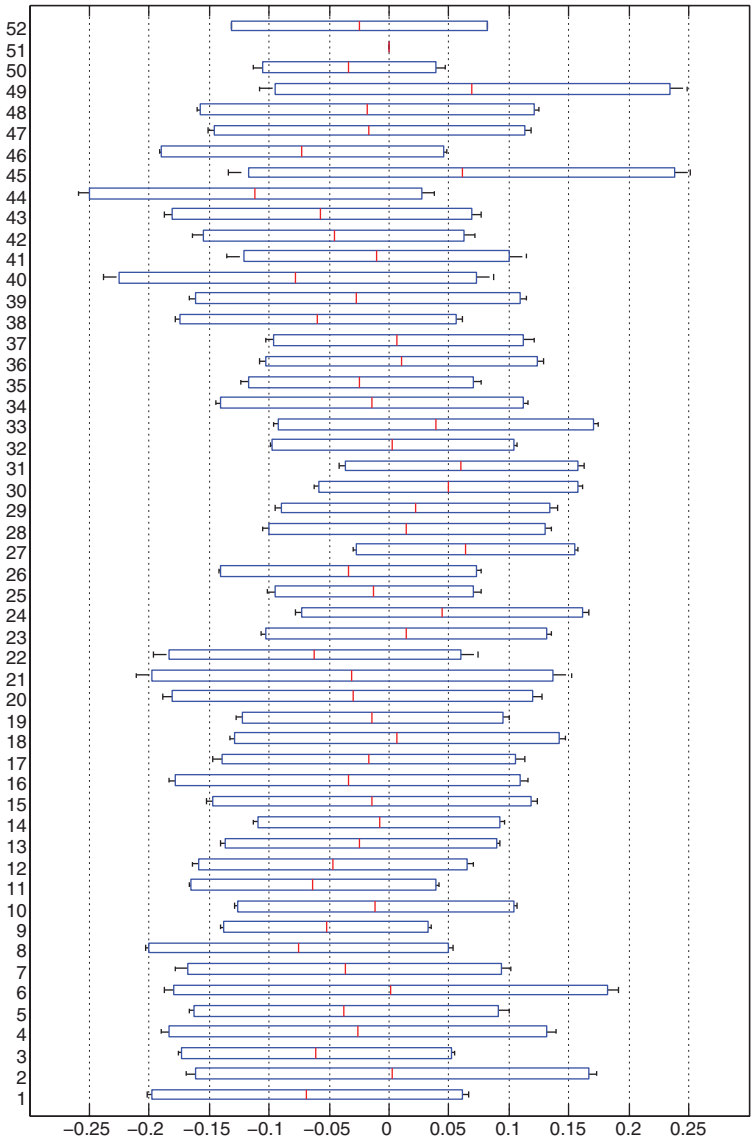

Figure 3: Bounds on ATE from US Election Example.
For each state, left and right of rectangular box corresponds to upper and lower bound of estimated ATE, as reported in Table 3.
The confidence sets are marked by the whiskers.
Y-axis: numbers correspond to states in Table 3.

$$\hat{G}_{W|D}(a|d)=\hat{F}_{W|D}\left(\frac{a}{\hat{p}_1}|d\right)=1-\hat{F}_{P|D}\left(\frac{\hat{p}_1}{a}|d\right),$$
$$\hat{G}_{V|D}(a|d)=\hat{F}_{V|D}\left(\frac{a}{1-\hat{p}_1}|d\right)=\hat{F}_{P|D}\left(1-\frac{1-\hat{p}_1}{a}|d\right),$$
$$\hat{G}_{V/W|D}(a|d)=\hat{F}_{V/W|D}\left(\frac{a\hat{p}_1}{1-\hat{p}_1}|d\right)=\hat{F}_{P|D}\left(\frac{a\hat{p}_1}{1-(1-a)\hat{p}_1}|d\right).$$

Noting that

$$G^{-1}_{W|D}(u|d)=\inf\left\{a:F_{W|D}\left(\frac{a}{p_1}|d\right)\geq u\right\}=p_1Q_{W|D}(u|d)$$

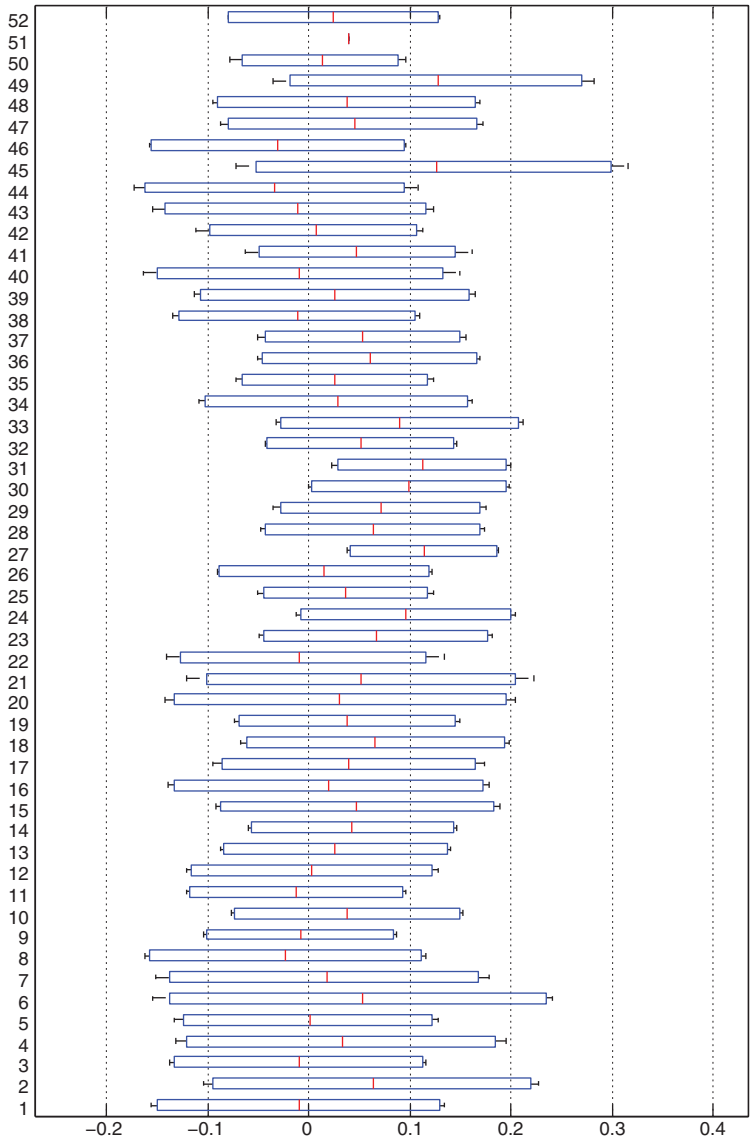

Figure 4: Bounds on ATT from US Election Example.
For each state, top and bottom of rectangular box corresponds to upper and lower bound of estimated ATT, as reported in Table 3.
The confidence sets are marked by the whiskers.
Y-axis: numbers correspond to states in Table 3.

and

$$G^{-1}_{V|D}(u|d)=\inf\left\{a: F_{V|D}\left(\frac{a}{1-p_1}|d\right)\geq u\right\}=(1-p_1)Q_{V|D}(u|d),$$

we have

$$\hat{\mu}^L_1=\int_0^{1-P_{01}}\hat{G}^{-1}_{W|D}(u|1)du,\quad \hat{\mu}^U_1=\int_{P_{01}}^1\hat{G}^{-1}_{W|D}(u|1)du,$$
$$\hat{\mu}^L_0=\int_0^{1-P_{00}}\hat{G}^{-1}_{V|D}(u|0)du,\quad \hat{\mu}^U_0=\int_{P_{00}}^1\hat{G}^{-1}_{V|D}(u|0)du.$$

Similarly,

$$\hat{\mu}^L_{0|1}=\int_0^{1-P_{00}}\hat{G}^{-1}_{V/W|D}(u|0)du,\quad \hat{\mu}^U_{0|1}=\int_{P_{00}}^1\hat{G}^{-1}_{V/W|D}(u|0)du.$$

The alternative expressions for the bounds estimators reveal that they are smooth functionals of $\hat{G}_{W|D}(\cdot|d)$, $\hat{G}_{V|D}(\cdot|d)$, and $\hat{G}_{V/W|D}(\cdot|d)$ which in turn are "empirical distribution functions" depending on the estimator of the propensity score $\hat{P}_i$ and $\hat{p}_d$. To characterize the effect of $\hat{P}_i$, we assume a semiparametric model for the propensity score $p(z)$, i.e. $p(z)=p(z;\beta_0,\tau_0)$ for some $\beta_0\in\mathcal{B}$ and $\tau_0\in\mathcal{T}$, where $\mathcal{B}$ is a compact subset of an Euclidean space and $\mathcal{T}$ is an infinite dimensional space (see Linton, Song, and Whang [2010]). This set-up subsumes the most commonly used parametric (such as probit or logit) and semiparametric models (including partially linear and single index models). Let $(\hat{\beta},\hat{\tau})$ denote consistent estimators of (β_0,τ_0) using dataset $\{Z_i, D_i\}_{i=1}^n$. Then $\hat{p}(z)\equiv p(z;\hat{\beta},\hat{\tau})$.

To avoid any confusion, we use p_{1o} to denote the true value of p_1. Let $\theta=(p_1,\beta)$ and $\theta_0=(p_{1o},\beta_0)$. Further define $B_{\mathcal{B}\times\mathcal{T}}(\delta)=\{(\beta,\tau)\in\mathcal{B}\times\mathcal{T}: \|\beta-\beta_0\|+\|\tau-\tau_0\|_\infty<\delta\}$ and $B_{\Theta\times\mathcal{T}}(\delta)=\{(\theta,\tau)\in\Theta\times\mathcal{T}: \|\theta-\theta_0\|+\|\tau-\tau_0\|_\infty<\delta\}$ for $\delta>0$, where $\|\cdot\|$ denotes the Euclidean norm and $\|\cdot\|_\infty$ denotes the sup-norm. Let $N(\varepsilon,\mathcal{T},\|\cdot\|_\infty)$ denote the ε-covering number of $\mathcal{T}$ with respect to $\|\cdot\|_\infty$. Let $\mathcal{P}$ be the collection of all the potential distributions of (Z_i, D_i) that satisfy Assumptions (s), (p) and (b) below.[10]

Assumption (s) (i) $p(Z_i)$ is a continuous random variable. (ii) There exists small positive constants ϵ_1, ϵ_2 such that $p(z)\in[\epsilon_1, 1-\epsilon_2]$ for all $z\in\mathcal{Z}$. (iii) Let $f_{P|D}(\cdot|d)$ denote the conditional pdf of $p(Z_i)$ given $D_i=d$. We assume $f_{P|D}(\cdot|d)$ is bounded away from zero on its support.

Assumption (p) (i) $\{(Z_i, D_i)\}_{i=1}^n$ is a random sample.

(ii) $\log N(\epsilon,\mathcal{T},\|\cdot\|_\infty)\le C\varepsilon^{-d_\tau}$ for some $d_\tau\in(0,1]$.

(iii) $f_{P|D}(\cdot|d)$ is bounded on its support and continuously differentiable with bounded derivatives on its support; and for some $\delta>0$, there exists a functional $\Gamma_p(\cdot|d)[\beta-\beta_0,\tau-\tau_0]$ of $(\beta-\beta_0,\tau-\tau_0)$, $(\beta,\tau)\in B_{\mathcal{B}\times\mathcal{T}}(\delta)$, such that for all $p\in[\epsilon_1,1-\epsilon_2]$,

$$|F_{P|D}(p;\beta,\tau|d)-F_{P|D}(p|d)-\Gamma_p(p|d)[\beta-\beta_0,\tau-\tau_0]|\le C_1\|\beta-\beta_0\|^2+C_2\|\tau-\tau_0\|_\infty^2; \tag{11}$$

and for $(\theta,\tau)\in B_{\Theta\times\mathcal{T}}(\delta)$, for all $w\in\mathcal{W}$,

$$\left|\Gamma_p\left(\frac{p_1}{w}|1\right)[\beta-\beta_0,\tau-\tau_0]-\Gamma_p\left(\frac{p_{1o}}{w}|1\right)[\beta-\beta_0,\tau-\tau_0]\right|\le C_1\|\theta-\theta_0\|^2+C_2|p_1-p_{1o}|\|\tau-\tau_0\|_\infty; \tag{12}$$

for all $v\in\mathcal{V}$,

$$\begin{aligned}&\left|\Gamma_p\left(1-\frac{1-p_1}{v}|0\right)[\beta-\beta_0,\tau-\tau_0]-\Gamma_p\left(1-\frac{1-p_{1o}}{v}|0\right)[\beta-\beta_0,\tau-\tau_0]\right|\\&\le C_1\|\theta-\theta_0\|^2+C_2|p_1-p_{1o}|\|\tau-\tau_0\|_\infty,\end{aligned} \tag{13}$$

with constants C_1 and C_2 that do not depend on P.

(iv) There exists δ, $C>0$ and a subvector Z_1 of Z such that: (a) the conditional density of Z given Z_1 is bounded uniformly over $(\beta,\tau)\in B_{\mathcal{B}\times\mathcal{T}}(\delta)$ and over $P\in\mathcal{P}$, (b) for each $(\beta,\tau)\in B_{\mathcal{B}\times\mathcal{T}}(\delta)$ and $(\beta',\tau')\in B_{\mathcal{B}\times\mathcal{T}}(\delta)$, $p(Z;\beta,\tau)-p(Z;\beta',\tau')$ is measurable with respect to the σ-field of Z_1, and (c) for each $(\beta_1,\tau_1)\in B_{\mathcal{B}\times\mathcal{T}}(\delta)$ and for each $\delta>0$,

$$\sup_{P\in\mathcal{P}}\sup_{z_1}E_P\left[\sup_{(\beta_2,\tau_2)\in B_{\mathcal{B}\times\mathcal{T}}(\delta)}|p(Z;\beta_1,\tau_1)-p(Z;\beta_2,\tau_2)|^2|Z_1=z_1\right]\le C\delta^{2s},$$

for some $s\in(d_\tau,1]$ with d_τ in Assumption (p) (ii), where the supremum over z_1 runs in the support of Z_1.

10 In Assumptions (s), (p), and (b), C, C_1, C_2 are generic constants which may take different values in different assumptions.

Assumption (b) (i) For each $\varepsilon>0$, $\sup_{P\in\mathcal{P}}P(\|\hat{\beta}-\beta_0\|+\|\hat{\tau}-\tau_0\|_\infty>\varepsilon)=o(1)$ and $\sup_{P\in\mathcal{P}}P(\hat{\tau}\in\mathcal{T})\to1$ as $n\to\infty$ such that $\|\hat{\beta}-\beta_0\|=o_p(n^{-1/4})$ and $\|\hat{\tau}-\tau_0\|_\infty=o_p(n^{-1/4})$ uniformly in $P\in\mathcal{P}$.

(ii) For each $\varepsilon>0$, uniformly $p\in[\epsilon_1, 1-\epsilon_2]$,

$$\sup_{P\in\mathcal{P}}P\left(\left|\sqrt{n}\Gamma_P(p|d)[\hat{\beta}-\beta_0, \hat{\tau}-\tau_0]-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i(p;\beta_0,\tau_0,d)\right|>\varepsilon\right)\to0;$$

where $\psi_i(p;\beta_0,\tau_0,d)$ depends on (Z_i, D_i) such that there exists $\eta>0$ such that for all $p\in[\epsilon_1, 1-\epsilon_2]$, $E_P[\psi_i(p;\beta_0,\tau_0,d)]=0$, and

$$\sup_{P\in\mathcal{P}}E_P\left[\sup_{w\in\mathcal{W}}\left|\psi_i(p;\beta_0,\tau_0,d)\right|^{2+\eta}\right]<\infty.$$

(iii) There exist constants $C>0$ and $s_1\in(d_\tau/2, 1)$ with d_τ in Assumption (p) (ii) such that for each $p_1\in[\epsilon_1, 1-\epsilon_2]$, and for each $\varepsilon>0$,

$$E\left[\sup_{p\in[\epsilon_1,1-\epsilon_2]:|p-p_1|\le\varepsilon}|\psi_i(p;\beta_0,\tau_0,d)-\psi_i(p_1;\beta_0,\tau_0,d)|^2\right]\le C\varepsilon^{2s_1}.$$

Theorem 6.1 *Suppose assumptions* (s), (p) *and* (b) *hold. Let* $\{\nu_W(\cdot), \nu_V(\cdot)\}$*be a bivariate Gaussian process with zero mean and covariance kernel given by* $C((w_1, v_1), (w_2, v_2))$ *defined as:*

$$C((w_1, v_1), (w_2, v_2))=\begin{pmatrix}E(\nu_W(w_1|1)\nu_W(w_2|1)) & E(\nu_W(w_1|1)\nu_V(v_2|1))\\ E(\nu_W(w_2|1)\nu_V(v_1|1)) & E(\nu_V(v_1|1)\nu_V(v_2|1))\end{pmatrix},$$

where

$$\begin{aligned}E(\nu_W(w_1|1)\nu_W(w_2|1))&=Cov(V_{i,W}(w_1;\theta_0,\tau_0), V_{i,W}(w_2;\theta_0,\tau_0)),\\ E(\nu_W(w_1|1)\nu_V(v_2|1))&=Cov(V_{i,W}(w_1;\theta_0,\tau_0), V_{i,V}(v_2;\theta_0,\tau_0)),\\ E(\nu_V(v_1|1)\nu_W(w_2|1))&=Cov(V_{i,V}(v_1;\theta_0,\tau_0), V_{i,W}(w_2;\theta_0,\tau_0)),\\ E(\nu_V(v_1|1)\nu_V(v_2|1))&=Cov(V_{i,V}(v_1;\theta_0,\tau_0), V_{i,V}(v_2;\theta_0,\tau_0)).\end{aligned}$$

in which

$$\begin{aligned}V_{i,W}(w;\theta_0,\tau_0)=&-\frac{1}{p_{1o}}[I\{p(Z_i;\beta_0,\tau_0)/p_{1o}\le1/w\}-[1-G_{W|D}(w|1)]]I\{D_i=1\}\\ &+\frac{1}{w}f_{P|D}(p_{1o}/w|1)[I\{D_i=1\}-p_{1o}]+\psi_i(p_{1o}/w;\beta_0,\tau_0,1)\end{aligned}\tag{14}$$

and

$$\begin{aligned}V_{i,V}(v;\theta_0,\tau_0)=&\frac{1}{(1-p_{1o})}\left[I\left\{p(Z_i;\beta_0,\tau_0)\le1-\frac{1-p_{1o}}{v}\right\}-G_{V|D}(v|0)\right]I\{D_i=0\}\\ &+\frac{1}{v(1-p_{1o})}f_{P|D}\left(1-\frac{1-p_{1o}}{v}|0\right)(I\{D_i=0\}-p_{1o})+\frac{1}{(1-p_{1o})}\psi_i\left(1-\frac{1-p_{1o}}{v};\beta_0,\tau_0,0\right).\end{aligned}\tag{15}$$

Then, uniformly in $P\in\mathcal{P}$,

$$\sqrt{n}\begin{pmatrix}\hat{\mu}_1^L-\mu_1^L\\ \hat{\mu}_1^U-\mu_1^U\\ \hat{\mu}_0^L-\mu_0^L\\ \hat{\mu}_0^U-\mu_0^U\end{pmatrix}\Rightarrow N(0,\Sigma_\mu),$$

where $\Sigma_\mu = E(Z_{10}Z_{10}')$ *in which* $Z_{10}=(Z_{1L}, Z_{1U}, Z_{0L}, Z_{0U})'$ *with*

$$Z_{1L}=\int_0^{1-P_{01}} \frac{\nu_W}{g_{W|D}} \circ G_{W|D}^{-1}(u|1)\,du,$$
$$Z_{1U}=\int_{P_{01}}^{1} \frac{\nu_W}{g_{W|D}} \circ G_{W|D}^{-1}(u|1)\,du,$$
$$Z_{0L}=\int_0^{1-P_{00}} \frac{\nu_V}{g_{V|D}} \circ G_{V|D}^{-1}(u|0)\,du,$$
$$Z_{0U}=\int_{P_{00}}^{1} \frac{\nu_V}{g_{V|D}} \circ G_{V|D}^{-1}(u|0)\,du.$$

Below we provide a sketch of the main steps in the proof of Theorem 6.1 and a discussion of the role of each of the Assumptions (s), (p) and (b) in the proof.

Step 1 We show that the random function,

$$(w, v) \mapsto \sqrt{n}\begin{pmatrix} \hat{G}_{W|D}(w|1)-G_{W|D}(w|1) \\ \hat{G}_{V|D}(v|0)-G_{V|D}(v|0) \end{pmatrix}$$

converges weakly to the Gaussian process $\{\nu_W(\cdot), \nu_V(\cdot)\}$ uniformly in $P\in\mathcal{P}$ and the convergence is in $D(\mathcal{W})\times D(\mathcal{V})$. This is done using empirical process methods similar to the proof of theorem 1 in Linton, Song, and Whang (2010); see also Andrews (1994), Chen, Linton, and van Keilegom (2003), and Linton, Maasoumi, and Whang (2005). Assumptions (p) and (b) are mainly used in this step. They are sufficient conditions for the Assumptions 1–3 needed for Lemma A.1 (in the Appendix).

Assumption p (ii) restricts the complexity of the class of functions that τ_0 belongs to. This and Assumption p (iv) ensure that the classes of functions:

$$\{I\{p(\cdot;\beta,\tau)\le p_1/w\}: w\in\mathcal{W}, (\theta,\tau)\in B_{\Theta\times\mathcal{T}}(\delta)\}$$

and

$$\{I\{p(\cdot;\beta,\tau)\le 1-(1-p_1)/v\}: v\in\mathcal{V}, (\theta,\tau)\in B_{\Theta\times\mathcal{T}}(\delta)\}$$

are uniformly Donsker so that the effect of estimating p_{1o} and (β_0, τ_0) can be analyzed via the behavior of $F_{P|D}(p_1/w;\beta,\tau|d)$ and $F_{P|D}(1-(1-p_1)/v;\beta,\tau|d)$. Eq. (11) in Assumption (p) (iii) implies that $F_{P|D}(p;\beta,\tau|d)$ can be approximated by linear functionals uniformly in p and Eqs. (12) and (13) in Assumption (p) (iii) impose further smoothness conditions on the relevant linear functionals. Together with Assumption (b) on uniform linear representations of these functionals evaluated at $(\hat{\beta}, \hat{\tau})$, they allow us to establish the stated weak convergence result in this step. The example below verifies these assumptions for a parametric model for the propensity score. Semiparametric models such as single-index models may be shown to satisfy these assumptions as well.

In general, suppose Z contains at least one component with an absolutely continuous distribution and $p(\cdot;\beta,\tau)$ is monotonically increasing in that component. Without loss of generality, let the first component of Z be that variable and the inverse of $p(\cdot;\beta,\tau)$ with respect to the first variable be $q(\cdot; Z_{-1};\beta,\tau)$. Then

$$\begin{aligned} F_{P|D}(p;\beta,\tau|d) &= \Pr(p(Z;\beta,\tau))(\le p|D=d) \\ &= \Pr(Z_1\le q(p, Z_{-1};\beta,\tau)|D=d) \\ &= E[F_{Z_1|D}(q(p, Z_{-1};\beta,\tau))|D=d]. \end{aligned}$$

Suppose $q(\cdot; Z_{-1}; \beta, \tau)$ satisfies: for $(\beta, \tau) \in B_{\mathcal{B}\times\mathcal{T}}(\delta)$,

$$\left|q(p, Z_{-1}; \beta, \tau) - q(p, Z_{-1}; \beta_0, \tau_0) - \Gamma_q(p, Z_{-1}|d)[\beta-\beta_0, \tau-\tau_0]\right| \le C_1\|\beta-\beta_0\|^2 + C_2\|\tau-\tau_0\|_\infty^2. \tag{16}$$

Let

$$\Gamma_p(p|d)[\beta-\beta_0, \tau-\tau_0] = E[f_{Z_1|D}(q(p, Z_{-1}; \beta, \tau))\Gamma_q(p, Z_{-1}|d)[\beta-\beta_0, \tau-\tau_0]]. \tag{17}$$

Then under some conditions, we can verify Assumption (p) (iii) and Assumption (b) (ii) for estimators of β, τ satisfying Assumption (b) (i). Crucial to Assumption (b) (ii) in the presence of τ is the existence of component Z_{-1} resulting in the smoothing operation in the expression for $\Gamma_p(p|d)[\beta-\beta_0, \tau-\tau_0]$ in (17).

Step 2 We show that for all constants c_1, c_2, c_3, c_4, the map $\phi_{F1, F0} : D(\mathcal{W})\times D(\mathcal{V}) \to R^2$ defined as

$$\phi_{F_1,F_0} = c_1\int_0^{1-P_{01}} F_1^{-1}(u)du + c_2\int_{P_{01}}^1 F_1^{-1}(u)du + c_3\int_0^{1-P_{00}} F_0^{-1}(u)du + c_4\int_{P_{00}}^1 F_0^{-1}(u)du$$

is Hadamard-differentiable at $(G_{W|D}(\cdot|1), G_{V|D}(\cdot|0))$ tangentially to $C(\mathcal{W})\times C(\mathcal{V})$ with the derivative

$$\begin{aligned}\phi'_{F_1,F_0}(\alpha_W, \alpha_V) \mapsto\ & c_1\int_0^{1-P_{01}} \frac{\alpha_W}{g_{W|D}} \circ G_{W|D}^{-1}(u|1)du + c_2\int_{P_{01}}^1 \frac{\alpha_W}{g_{W|D}} \circ G_{W|D}^{-1}(u|1)du \\ & + c_3\int_0^{1-P_{00}} \frac{\alpha_V}{g_{V|D}} \circ G_{V|D}^{-1}(u|0)du + c_4\int_{P_{00}}^1 \frac{\alpha_V}{g_{V|D}} \circ G_{V|D}^{-1}(u|0)du,\end{aligned}$$

where $g_{w|D}$ and $g_{V|D}$ are pdfs corresponding to $G_{W|D}$ and $G_{V|D}$. Assumption (s) is used in this step. It ensures that the quantile functions (F_1^{-1}, F_0^{-1}) are Hadamard-differentiable at $(G_{W|D}(\cdot|1), G_{V|D}(\cdot|0))$ tangentially to $C(\mathcal{W})\times C(\mathcal{V})$; see van der Vaart and Wellner (1996).[11] By the Functional Delta method (see van der Vaart and Wellner [1996]), we obtain that, uniformly in $P \in \mathcal{P}$,

$$\sqrt{n}(c_1\hat{\mu}_1^L + c_2\hat{\mu}_1^U + c_3\hat{\mu}_0^L + c_4\hat{\mu}_0^U - [c_1\mu_1^L + c_2\mu_1^U + c_3\mu_0^L + c_4\mu_0^U]) \Rightarrow \phi'_{F_1,F_0}(\nu_W, \nu_V).$$

Theorem 6.1 can be used to construct confidence sets (CSs) for the mean outcomes μ_1, μ_0. It also allows us to construct CSs for the ATE μ_Δ, as an application of the Delta method leads to the asymptotic distributions of the estimators of the lower and upper bounds on: $\mu_\Delta : (\hat{\mu}_1^L - \hat{\mu}_0^U, \hat{\mu}_1^U - \hat{\mu}_0^L)$.

The second term in the expression for $V_{i,W}(w; \theta_0, \tau_0)$ $(V_{i,V}(v; \theta_0, \tau_0))$ in (14) (15) accounts for the effect of estimating p_{1o} by $\hat{p}_1$ and the third term accounts for the effect of estimating (β_0, τ_0) by $(\hat{\beta}, \hat{\tau})$; these take the more familiar form when the model for the propensity score is parametric as illustrated below.

Example: A parametric model for the propensity score Let $p(z) = p(z; \beta_0)$ for some $\beta_0 \in \mathcal{B}$. Such a parametric model for the propensity score was considered in the empirical application above. If $F_{P|D}(p; \beta|d)$ is twice differentiable in (p, β) with bounded second order derivatives in p and a neighborhood of β_0, Assumption (p) (iii) is satisfied with

$$\Gamma_p(p|d)[\beta-\beta_0] = \frac{\partial F_{P|D}(p; \beta_0|d)}{\partial \beta'}(\beta-\beta_0)$$

and

11 We note that Assumption (s) may be relaxed at the expense of a more tedious proof analogous to the proof of claim 1 in Bhattacharya (2007), see also Goldie (1977).

$$
\begin{aligned}
&\left|\Gamma_P\left(\frac{p_1}{w}|1\right)[\beta-\beta_0]-\Gamma_P\left(\frac{p_{1o}}{w}|1\right)[\beta-\beta_0]\right| \\
&=\left|\left[\frac{\partial F_{P|D}\left(\frac{p_1}{w};\beta_0|1\right)}{\partial\beta'}-\frac{\partial F_{P|D}\left(\frac{p_{1o}}{w};\beta_0|1\right)}{\partial\beta'}\right](\beta-\beta_0)\right| \\
&\le C_1\|\theta-\theta_0\|^2, \\
&\left|\Gamma_P\left(1-\frac{1-p_1}{v}|0\right)[\beta-\beta_0]-\Gamma_P\left(1-\frac{1-p_{1o}}{v}|0\right)[\beta-\beta_0]\right| \\
&=\left|\left[\frac{\partial F_{P|D}\left(1-\frac{1-p_1}{v};\beta_0|0\right)}{\partial\beta'}-\frac{\partial F_{P|D}\left(1-\frac{1-p_{1o}}{v};\beta_0|0\right)}{\partial\beta'}\right](\beta-\beta_0)\right| \\
&\le C_1\|\theta-\theta_0\|^2.
\end{aligned}
$$

Assumption (p) (iv) is satisfied if $p(z;\beta)$ is twice differentiable with bounded second of β_0. Let $\hat{\beta}$ satisfy:

$$\sqrt{n}(\hat{\beta}-\beta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_i(\beta_0)+o_P(1)$$

where $\varphi_i(\beta_0)$ satisfies: $E(\varphi_i(\beta_0))=0$ and $E(\varphi_i^{2+\eta}(\beta_0))<\infty$ for some $\eta>0$. Then Assumption (b) is satisfied with

$$\psi_i(p;\beta_0,d)=\frac{\partial F_{P|D}(p;\beta_0|d)}{\partial\beta'}\varphi_i(\beta_0).$$

To establish the asymptotic distribution of $(\hat{\mu}_{0|1}^L,\hat{\mu}_{0|1}^U)$, we replace Assumption (p) (iii) with Assumption (p) (iii)′ below.

Assumption (p) (iii)′: $f_{P|D}(\cdot|d)$ is bounded on its support and continuously differentiable with bounded derivatives on its support; and for some $\delta>0$, there exists a functional $\Gamma_P(\cdot|d)[\beta-\beta_0,\tau-\tau_0]$ of $(\beta-\beta_0,\tau-\tau_0)$, $(\beta,\tau)\in B_{B\times T}(\delta)$, such that for all $p\in[\epsilon_1,1-\epsilon_2]$, (11) holds, and for $(\theta,\tau)\in B_{\Theta\times T}(\delta)$, all $a\in\mathcal{A}$,

$$
\begin{aligned}
&\left|\Gamma_P\left(\frac{ap_1}{1-(1-a)p_1}|0\right)\left[\beta-\beta_0,\tau-\tau_0\right]-\Gamma_P\left(\frac{ap_{1o}}{1-(1-a)p_{1o}}|0\right)\left[\beta-\beta_0,\tau-\tau_0\right]\right| \\
&\le C_1\left\|\theta-\theta_0\right\|^2+C_2\left|p_1-p_{1o}\right|\left\|\tau-\tau_0\right\|_\infty,
\end{aligned}
\tag{18}
$$

with constants C_1 and C_2 that do not depend on P.

The following result (which is a special case of Theorem 6.1 above) obtains:

Theorem 6.2 *Let $\nu_{V|W}(\cdot)$ be a Gaussian process with zero mean and covariance kernel:*

$$E(V_{i,V/W}(a_1;\theta_0,\tau_0)V_{i,V/W}(a_2;\theta_0,\tau_0)),\ \textit{where}$$

$$
\begin{aligned}
&V_{i,V/W}(a;\theta_0,\tau_0) \\
&=\frac{1}{(1-p_{1o})}\left[I\left\{p(Z_i;\beta_0)\le\frac{ap_{1o}}{1-(1-a)p_{1o}}\right\}-G_{V/W|D}(v|0)\right]I\{D_i=0\} \\
&+\frac{1}{(1-p_{1o})}\psi_i\left(\frac{p_{1o}a}{1-(1-a)p_{1o}};\beta_0,\tau_0,d\right) \\
&+\frac{a}{(1-p_{1o})\left[1-(1-a)p_{1o}\right]^2}f_{P|D}\left(\frac{p_{1o}a}{1-(1-a)p_{1o}};\beta_0,\tau_0|0\right)(I\{D_i=1\}-p_{1o}).
\end{aligned}
$$

Suppose assumptions (*s*), (*p*) *with* (*p*) (*iii*) *replaced with* (*p*) (*iii*)′ *and* (*b*) *hold. Then uniformly in* $P\in\mathcal{P}$,

$$\sqrt{n}\begin{pmatrix}\hat{\mu}^L_{0|1}-\mu^L_{0|1}\\ \hat{\mu}^U_{0|1}-\mu^U_{0|1}\end{pmatrix}'\Rightarrow N(0,\Sigma_{\mu|1}),$$

where $\Sigma_{\mu|1}=E(Z'_{10|1}Z_{10|1})$ *in which* $Z_{10|1}=(Z^L_{0|1},Z^U_{0|1})'$ *with*

$$Z^L_{0|1}=\int_0^{1-P_{00}}\frac{\nu_{V/W}}{g_{V/W|D}}\circ G^{-1}_{V/W|D}(u|0)du,$$
$$Z^U_{0|1}=\int_{P_{00}}^1\frac{\nu_{V/W}}{g_{V/W|D}}\circ G^{-1}_{V/W|D}(u|0)du.$$

Theorem 6.2 allows us to construct CSs for the counterfactual mean outcome $\mu_{0|1}$ and the treatment effect for the treated $\mu_{\Delta|1}$, as $\mu_{\Delta|1}=\mu_{1|1}-\mu_{0|1}$ and $\hat{\mu}_{1|1}=1-P_{01}$ is treated as a constant.

7 Conclusion

Combining tools and insights from the treatment effect and copula literatures, this paper has presented a novel approach to counterfactual analysis in EI models, in which aggregate and individual-level data must be combined in order to infer individual-level behavior. Under a "selection on observables" assumption familiar from the treatment effects literature, we establish partial identification results for the mean and other functionals of the counterfactual outcome distribution.

We provide estimators and derive inference tools by using empirical process methods combined with recent developments on inference for partially identified parameters. Inference tools for distributional treatment effect parameters such as $F_\Delta(\delta|D=1)$ may be established by generalizing the technical tools in Fan and Song (2011). This is challenging and left for future research.

Acknowledgments: We are grateful to Cheng Hsiao, Sergio Firpo, Chuck Manski, Kevin Song, and Jeff Wooldridge for valuable comments and discussions. We thank SangMok Lee for excellent research assistance, and seminar participants at Michigan State, USC, and the Canadian Econometrics Study Group meetings (2011, Toronto) for useful comments.

Appendix

Appendix: Technical Proofs

The proofs of Theorems 6.1 and 6.2 are similar; they rely heavily on the lemma below which is adapted from the proof of Theorem 1 in Linton, Song, and Whang (2010). Closely related work include Andrews (1994), Chen, Linton, and van Keilegom (2003), and Linton, Maasoumi, and Whang (2005).

Let

$$X_i(\theta,\tau)=\varphi(Z_i;\theta,\tau),$$

where $\varphi(\cdot;\theta,\tau)$ is a real valued function known up to the parameter $(\theta,\tau)\in\Theta\times\mathcal{T}$ with Θ a compact subset of a Euclidean space and $\mathcal{T}$ an infinite dimensional space. For $d=1,0$, let $\nu_n(\cdot;d)$ be the stochastic process on $\mathcal{X}$ with

$$\nu_n(x;d)=\sqrt{n}\left[n^{-1}\sum_{i=1}^n I\{X_i(\hat{\theta},\hat{\tau})\le x\}I\{D_i=d\}-E(I\{X_i(\theta_0,\tau_0)\le x\}I\{D_i=d\})\right],$$

Where $x\in\mathcal{X}$, $(\theta_0, \tau_0)\in\Theta\times\mathcal{T}$, and $(\hat{\theta}, \hat{\tau})$ are consistent estimators of (θ_0, τ_0).

Lemma A.1 below presents conditions under which the process $\{\nu_n(\cdot; d)\}$ converges weakly to a Gaussian process.

Let $B_{\Theta\times\mathcal{T}}(\delta)=\{\theta, \tau\}\in\Theta\times\mathcal{T}$: $\|\theta-\theta_0\|+\|\tau-\tau_0\|_\infty<\delta\}$ for $\delta>0$ and $\mathcal{P}$ be the collection of all the potential distributions of (Z_i, D_i) that satisfy Assumptions 1–3 below.

Assumption 1 (i) $\{Z_i, D_i\}_{i=1}^n$ is a random sample.

(ii) $\log N(\varepsilon, \mathcal{T}, \|\cdot\|_\infty)\le C\varepsilon^{-d}$ for some $d\in(0, 1]$.

(iii) Let

$$F_{X|D}(x; \theta, \tau|d)=\Pr(X_i(\theta, \tau)\le x|D_i=d).$$

For some $\delta>0$, there exists a functional $\Gamma_{F,P}(x|d)[\theta-\theta_0, \tau-\tau_0]$ of $(\theta-\theta_0, \tau-\tau_0)$, $(\theta, \tau)\in B_{\Theta\times\mathcal{T}}(\delta)$ such that

$$\begin{aligned}&|F_{X|D}(x; \theta, \tau|d)-F_{X|D}(x; \theta_0, \tau_0|d)-\Gamma_{F,P}(x|d)[\theta-\theta_0, \tau-\tau_0]|\\&\le C_1\|\theta-\theta_0\|^2+C_2\|\tau-\tau_0\|_\infty^2,\end{aligned}$$

with constants C_1 and C_2 that do not depend on P.

Assumption 2 (i) $X_i(\theta_0, \tau_0)$ is a continuous random variable with a bounded support $\mathcal{X}$.

(ii) There exists δ, $C>0$ and a subvector Z_1 of Z such that: (a) the conditional density of Z given Z_1 is bounded uniformly over $(\theta, \tau)\in B_{\Theta\times\mathcal{T}}(\delta)$ and over $P\in\mathcal{P}$, (b) for each $(\theta, \tau)\in B_{\Theta\times\mathcal{T}}(\delta)$ and $(\theta', \tau')\in B_{\Theta\times\mathcal{T}}(\delta)$, $\varphi(Z; \theta, \tau)-\varphi(Z; \theta', \tau')$ is measurable with respect to the σ-field of Z_1, and (c) for each $(\theta_1, \tau_1)\in B_{\Theta\times\mathcal{T}}(\delta)$ and for each $\delta>0$,

$$\sup_{P\in\mathcal{P}}\sup_{z_1}E_P\left[\sup_{(\theta_2,\tau_2)\in B_{\Theta\times\mathcal{T}}(\delta)}|\varphi(Z; \theta_1, \tau_1)-\varphi(Z; \theta_2, \tau_2)|^2|Z_1=z_1\right]\le C\delta^{2s},$$

for some $s\in(d, 1]$ with d in Assumption 1 (ii), where the supremum over z_1 runs in the support of Z_1.

Assumption 3 (i) For each $\varepsilon>0$, $\sup_{P\in\mathcal{P}}P(\|\hat{\theta}-\theta_0\|+\|\hat{\tau}-\tau_0\|_\infty>\varepsilon)=o(1)$ and $\sup_{P\in\mathcal{P}}P(\hat{\tau}\in\mathcal{T})\to1$ as $n\to\infty$ such that $\|\hat{\theta}-\theta_0\|=o_P(n^{-1/4})$ and $\|\hat{\tau}-\tau_0\|_\infty=o_P(n^{-1/4})$.

(ii) For each $\varepsilon>0$,

$$\sup_{P\in\mathcal{P}}P\left(\left|\sqrt{n}\Gamma_{F,P}(x|d)[\hat{\theta}-\theta_0, \hat{\tau}-\tau_0]-\frac{1}{\sqrt{n}}\sum_{i=1}^n\psi_{x,F}(Z_i, D_i; \theta_0, \tau_0, d)\right|>\varepsilon\right)\to0,$$

where $\psi_{x,F}(Z_i, D_i, \theta_0, \tau_0, d)$ satisfies that there exists $\eta>0$ such that for all $x\in\mathcal{X}$ $E_P[\psi_{x,F}(Z_i, D_i, \theta_0, \tau_0, d)]=0$ and

$$\sup_{P\in\mathcal{P}}E_P\left[\sup_{x\in\mathcal{X}}|\psi_{x,F}(Z_i, D_i; \theta_0, \tau_0, d)|^{2+\eta}\right]<\infty.$$

(iii) There exist constants $C>0$ and $s_1\in(d/2,1]$ with d in Assumption 1 (ii) such that for each $x_1\in\mathcal{X}$ and for each $\varepsilon>0$,

$$E\left[\sup_{x\in\mathcal{X}:|x-x_1|\le\varepsilon}|\psi_{x,F}(Z_i, D_i; \theta_0, \tau_0, d)-\psi_{x_1,F}(Z_i, D_i; \theta_0, \tau_0, d)|^2\right]\le C\varepsilon^{2s_1}.$$

Let $\nu(\cdot; d)$ be a mean zero Gaussian process on $\mathcal{X}$ with a covariance kernel given by

$$C(x_1, x_2; d)=Cov(V_i(x_1; \theta_0, \tau_0, d), V_i(x_2; \theta_0, \tau_0, d)),$$

where

$$V_i(x;\theta_0,\tau_0,d)=I\{X_i(\theta_0,\tau_0)\le x\}I\{D_i=1\}+\psi_{x,F}(Z_i,D_i;\theta_0,\tau_0,d).$$

Lemma A.1 *Suppose that Assumptions 1–3 hold. Then*

$$\nu_n(x;d)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[V_i(x;\theta_0,\tau_0,d)-E(V_i(x;\theta_0,\tau_0,d))]+o_P(1)$$

uniformly in $x\in\mathcal{X}$ and $P\in\mathcal{P}$ and hence $\nu_n(\cdot;d)$ weakly converges to $\nu(\cdot;d)$ uniformly in $P\in\mathcal{P}$.

Proof of Theorem 6.1: We will show that for any constants c_1, c_2, c_3, c_4, the linear combination $c_1\hat{\mu}_1^L+c_2\hat{\mu}_1^U+c_3\hat{\mu}_0^L+c_4\hat{\mu}_0^U$ is asymptotically normally distributed with variance $(c_1,c_2,c_3,c_4)\Sigma_\mu(c_1,c_2,c_3,c_4)'$.

Assumption (s) ensures that $G_{W|D}(\cdot|1)G_{V|D}(\cdot|0)$ have compact supports and the corresponding pdfs are bounded away from zero on their supports. As a result , the map $\phi_{F1,F0}:D(\mathcal{W})\times D(\mathcal{V})\to R$ defined as

$$\begin{aligned}\phi_{F_1,F_0}=&c_1\int_0^{1-P_{01}}F_1^{-1}(u)du+c_2\int_{P_{01}}^{1}F_1^{-1}(u)du\\&+c_3\int_0^{1-P_{00}}F_0^{-1}(u)du+c_4\int_{P_{00}}^{1}F_0^{-1}(u)du\end{aligned}$$

is Hadamard-differentiable at $(G_{W|D}(\cdot|1),\ G_{V|D}(\cdot|0))$ tangentially to $C(\mathcal{W})\times C(\mathcal{V})$ with the derivative:

$$\begin{aligned}\phi'_{F_1,F_0}(\alpha_W,\alpha_V)\mapsto&c_1\int_0^{1-P_{01}}\frac{\alpha_W}{g_{W|D}}\circ G_{W|D}^{-1}(u|1)du+c_2\int_{P_{01}}^{1}\frac{\alpha_W}{g_{W|D}}\circ G_{W|D}^{-1}(u|1)du\\&+c_3\int_0^{1-P_{00}}\frac{\alpha_V}{g_{V|D}}\circ G_{V|D}^{-1}(u|0)du+c_4\int_{P_{00}}^{1}\frac{\alpha_V}{g_{V|D}}\circ G_{V|D}^{-1}(u|0)du,\end{aligned}$$

see van der Vaart and Wellner (1996).

We will complete the proof by establishing the weak convergence of the stochastic process:

$$\{\sqrt{n}(\hat{G}_{W|D}(w|1)-G_{W|D}(w|1),\ \hat{G}_{V|D}(v|0)-G_{V|D}(v|0))':(w,v)\in\mathcal{W}\times\mathcal{V}\}$$

and invoking the Functional Delta method.

Let

$$\begin{aligned}\nu_{nG,W}(w)&=\sqrt{n}[\hat{G}_{W|D}(w|1)\text{-}G_{W|D}(w|1)],\ w\in\mathcal{W},\\ \nu_{nG,V}(v)&=\sqrt{n}[\hat{G}_{V|D}(v|0)\text{-}G_{V|D}(v|0)],\ v\in\mathcal{V}.\end{aligned}$$

Step 1 We show: $\nu_{nG,W}(w)=n^{-1/2}\sum_{j=1}^{n}V_{j,W}(w;\theta_0,\tau_0)+o_P(1)$ uniformly in $w\in\mathcal{W}$ and $P\in\mathcal{P}$.

By the definition of $\hat{G}_{W|D}$, we have:

$$\begin{aligned}\nu_{nG,W}(w)=&\sqrt{n}\left[1-\frac{n^{-1}\sum_{j=1}^{n}I\{p(Z_j;\hat{\beta},\hat{\tau})\le\hat{p}_1/w\}I\{D_j=1\}}{\hat{p}_1}-G_{W|D}(w|1)\right]\\=&-\frac{n^{-1/2}\sum_{j=1}^{n}I\{p(Z_j;\hat{\beta},\hat{\tau})/\hat{p}_1\le 1/w\}I\{D_j=1\}-p_1[1-G_{W|D}(w|1)]}{\hat{p}_1}\\&+\frac{\sqrt{n}(\hat{p}_1-p_1)}{\hat{p}_1}[1-G_{W|D}(w|1)]\\=&-\frac{n^{-1/2}\sum_{j=1}^{n}I\{p(Z_j;\hat{\beta},\hat{\tau})/\hat{p}_1\le 1/w\}I\{D_j=1\}-p_1[1-G_{W|D}(w|1)]}{p_1}\\&+\frac{\sqrt{n}(\hat{p}_1-p_1)}{p_1}[1-G_{W|D}(w|1)]+o_P(1).\end{aligned}$$

We apply Lemma A.1 to the first term on the right hand side of the last equation with $X_j(\theta, \tau)=p(Z_j; \beta, \tau)/p_1$ and $\theta=(p_1, \beta)$. We verify Assumptions 1–3 in Lemma A.1 under Assumptions (s), (p) and (b).

Assumption 1 (i) and (ii) hold under Assumption (p) (i) and (ii). Now we verify Assumption 1 (iii). Note that for $x=1/w$,

$$\begin{aligned}F_{X|D}(x; \theta, \tau|1)&=E_P[I\{X_i(\theta, \tau)\le x\}|D_i=1]\\&=E_P[I\{p(Z_i; \beta, \tau)\le p_1 x\}|D_i=1]\\&=F_{P|D}(p_1 x; \beta, \tau|1).\end{aligned}$$

Let

$$\Gamma_{F,P}(x|1)[\theta-\theta_0, \tau-\tau_0]=xf_{P|D}(p_{1o}x|1)(p_1-p_{1o})+\Gamma_P(p_{1o}x|1)[\beta-\beta_0, \tau-\tau_0],$$

where $\Gamma_P(p_{1o}x|1)[\beta-\beta_0, \tau-\tau_0]$ is defined in Assumption (p) (iii). Then by Assumption (p) (iii), we conclude: for some $\delta>0$, $(\theta, \tau)\in B_{\Theta\times T}(\delta)$,

$$\begin{aligned}&|F_{X|D}(x; \theta, \tau|1)-F_{X|D}(x; \theta_0, \tau_0|1)-\Gamma_{F,P}(x|1)[\theta-\theta_0, \tau-\tau_0]|\\&=\left|\begin{array}{c}F_{P|D}(p_1 x; \beta, \tau|1)-F_{P|D}(p_{1o}x|1)\\-xf_{P|D}(p_{1o}x|1)(p_1-p_{1o})-\Gamma_P(p_{1o}x|1)[\beta-\beta_0, \tau-\tau_0]\end{array}\right|\\&=\left|\begin{array}{l}F_{P|D}(p_1 x; \beta, \tau|1)-F_{P|D}(p_1 x|1)-\Gamma_P(p_1 x|1)[\beta-\beta_0, \tau-\tau_0]\\+\Gamma_P(p_1 x|1)[\beta-\beta_0, \tau-\tau_0]-\Gamma_P(p_{1o}x|1)[\beta-\beta_0, \tau-\tau_0]\\+[F_{P|D}(p_1 x|1)-F_{P|D}(p_{1o}x|1)-xf_{P|D}(p_{1o}x|1)(p_1-p_{1o})]\end{array}\right|\\&\le|F_{P|D}(p_1 x; \beta, \tau|1)-F_{P|D}(p_1 x|1)-\Gamma_P(p_1 x|1)[\beta-\beta_0, \tau-\tau_0]|\\&\quad+\left|\Gamma_P(p_1 x|1)[\beta-\beta_0, \tau-\tau_0]-\Gamma_P(p_{1o}x|1)[\beta-\beta_0, \tau-\tau_0]\right|\\&\quad+\frac{1}{2}\left|x^2\frac{\partial^2}{\partial x^2}F_{P|D}(p_1^* x|1)(p_1-p_{1o})^2\right|\\&\le C_1\|\theta-\theta_0\|^2+C_2\|\tau-\tau_0\|_\infty^2,\end{aligned}$$

where p_1^* lies between p_{1o} and p_1.

Assumption 2 holds under Assumption (s) (i) and Assumption (p) (iv).

It remains to verify Assumption 3. Assumption 3 (i) holds because of Assumption (b) (i). For Assumption 3 (ii), we let

$$\psi_{x,F}(Z_i, D_i; \theta_0, \tau_0, 1)=xf_{P|D}(p_{1o}x|1)(I\{D_i=1\}-p_{1o})+\psi_i(p_{1o}x; \beta_0, \tau_0, 1),$$

where $\psi_i(p_{1o}x; \beta_0, \tau_0, 1)$ is defined in Assumption (b) (ii). Then by Assumption (b) (ii), we obtain:

$$\begin{aligned}&\sup_{P\in\mathcal{P}}P\left(\left|\sqrt{n}\Gamma_{F,P}(x|1)[\hat\theta-\theta_0, \hat\tau-\tau_0]-\frac{1}{\sqrt{n}}\sum_{i=1}^n\psi_{x,F}(Z_i, D_i; \theta_0, \tau_0, 1)\right|>\varepsilon\right)\\&=\sup_{P\in\mathcal{P}}P\left(\left|\sqrt{n}\Gamma_P(p_{1o}x|1)[\hat\beta-\beta_0, \hat\tau-\tau_0]-\frac{1}{\sqrt{n}}\sum_{i=1}^n\psi_i(p_{1o}x; \beta_0, \tau_0, 1)\right|>\varepsilon\right)\to 0,\end{aligned}$$

where $E_P[\psi_{x,F}(Z_i, D_i; \theta_0, \tau_0, 1)]=0$ and

$$\begin{aligned}&\sup_{P\in\mathcal{P}}E_P\left[\sup_{x\in\mathcal{X}}|\psi_{x,F}(Z_i, D_i; \theta_0, \tau_0, 1)|^{2+\eta}\right]\\&=\sup_{P\in\mathcal{P}}E_P\left[\sup_{x\in\mathcal{X}}|xf_{P|D}(p_{1o}x|1)(I\{D_i=1\}-p_{1o})+\psi_i(p_{1o}x; \beta_0, \tau_0, 1)|^{2+\eta}\right]\\&<\infty\end{aligned}$$

by Assumption (p) (iii) and Assumption (b) (ii). It remains to verify Assumption 3 (iii):

$$
\begin{aligned}
&E\left[\sup_{x\in\mathcal{X}:|x-x_1|\le\varepsilon}\left|\psi_{x,F}(Z_i,D_i;\theta_0,\tau_0,1)-\psi_{x_1,F}(Z_i,D_i;\theta_0,\tau_0,1)\right|^2\right]\\
&=E\left[\sup_{x\in\mathcal{X}:|x-x_1|\le\varepsilon}\left|\begin{array}{l} xf_{P|D}(p_{1o}x|1)(I\{D_i=1\}-p_{1o})+\psi_i(p_{1o}x;\beta_0,\tau_0,1)\\ -x_1f_{P|D}(p_{1o}x_1|1)(I\{D_i=1\}-p_{1o})-\psi_i(p_{1o}x;\beta_0,\tau_0,1)\end{array}\right|^2\right]\\
&\le CE\left[\sup_{x\in\mathcal{X}:|x-x_1|\le\varepsilon}|x-x_1|^2\right]\\
&+CE\left[\sup_{x\in\mathcal{X}:|x-x_1|\le\varepsilon}\left|\psi_i(p_{1o}x;\beta_0,\tau_0,d)-\psi_i(p_{1o}x_1;\beta_0,\tau_0,d)\right|^2\right]\\
&\le C\varepsilon^2+C\varepsilon^{2s_1}
\end{aligned}
$$

by Assumption (b) (iii) and Assumption (p) (iii).

Using Lemma A.1, we now obtain:

$$
\begin{aligned}
&\nu_{nG,W}(w)\\
&=-\frac{1}{p_{1o}}n^{-1/2}\sum_{j=1}^{n}\Big[I\{p(Z_j;\beta_0)/p_{1o}\le 1/w\}I\{D_j=1\}-p_{1o}[1-G_{W|D}(w|1)]\Big]\\
&-\frac{1}{w}f_{P|D}(p_{1o}/w;\beta_0,\tau_0|d)(\hat{p}_1-p_{1o})-\Gamma_P(p_{1o}/w|d)[\hat{\beta}-\beta_0,\hat{\tau}-\tau_0]\\
&+\frac{1}{p_{1o}}n^{-1/2}\sum_{j=1}^{n}[I\{D_j=1\}-p_{1o}][1-G_{W|D}(w|1)]+o_P(1)\\
&=-n^{-1/2}\sum_{j=1}^{n}\left\{\begin{array}{c}\frac{1}{p_{1o}}[I\{p(Z_j;\beta_0,\tau_0)/p_{1o}\le 1/w\}-[1-G_{W|D}(w|1)]]I\{D_j=1\}\\ +\frac{1}{w}f_{P|D}(p_{1o}/w;\beta_0,\tau_0|d)[I\{D_j=1\}-p_{1o}]\\ +\psi_j(p_{1o}/w;\beta_0,\tau_0,1)\end{array}\right\}\\
&+o_P(1)\\
&=n^{-1/2}\sum_{j=1}^{n}V_{j,W}(w;\theta_0,\tau_0)+o_P(1).
\end{aligned}
$$

Step 2 We show: $\nu_{nG,V}(v)=n^{-1/2}\sum_{j=1}^{n}V_{j,V}(v;\theta_0,\tau_0)+o_P(1)$..

Assumptions 1–3 in Lemma A.1 can be verified by following Step 1. So we just provide the main expressions. Note that $\hat{G}_{V|D}(v|d)=\hat{F}_{P|D}\left(1-\frac{1-\hat{p}_1}{v}|d\right)$ and

$$
E_P\left[I\left\{p(Z_i;\beta,\tau)\le 1-\frac{1-p_1}{v}\right\}|D_i=0\right]=F_{P|D}\left(1-\frac{1-p_1}{v};\beta,\tau|0\right).
$$

We let

$$
\begin{aligned}
&\Gamma_{F,P}(x|d)[\theta-\theta_0,\tau-\tau_0]\\
&=\frac{1}{v}f_{P|D}\left(1-\frac{1-p_{1o}}{v}|0\right)(p_1-p_{1o})+\Gamma_P\left(1-\frac{1-p_{1o}}{v}|0\right)[\beta-\beta_0,\tau-\tau_0].
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\nu_{nG,V}(v)\\
&=\sqrt{n}[\hat{G}_{V|D}(v|0)-G_{V|D}(v|0)]\\
&=\sqrt{n}\left[\hat{F}_{P|D}\left(1-\frac{1-\hat{p}_1}{v}|0\right)-G_{V|D}(v|0)\right]\\
&=n^{-1/2}\sum_{j=1}^{n}\left[\frac{1}{\hat{p}_0}I\left\{p(Z_j;\beta_0,\tau_0)\le 1-\frac{1-\hat{p}_1}{v}\right\}I\{D_j=0\}-G_{V|D}(v|0)\right]\\
&=n^{-1/2}\sum_{j=1}^{n}\left[I\left\{p(Z_j;\beta_0,\tau_0)\le 1-\frac{1-\hat{p}_1}{v}\right\}I\{D_j=0\}-p_0G_{V|D}(v|0)\right]\frac{1}{\hat{p}_0}\\
&\quad+\frac{\sqrt{n}(p_0-\hat{p}_0)}{\hat{p}_0}G_{V|D}(v|0)\\
&=n^{-1/2}\sum_{j=1}^{n}\left[I\left\{p(Z_j;\beta_0,\tau_0)\le 1-\frac{1-\hat{p}_1}{v}\right\}I\{D_j=0\}-p_0G_{V|D}(v|0)\right]\frac{1}{p_0}\\
&\quad+\frac{\sqrt{n}(p_0-\hat{p}_0)}{p_0}G_{V|D}(v|0)+o_P(1)\\
&=n^{-1/2}\sum_{j=1}^{n}\frac{1}{(1-p_{1o})}\left[I\left\{p(Z_j;\beta_0,\tau_0)\le 1-\frac{1-p_{1o}}{v}\right\}-G_{V|D}(v|0)\right]I\{D_j=0\}\\
&\quad+n^{-1/2}\sum_{j=1}^{n}\frac{1}{(1-p_{1o})}\psi_j\left(1-\frac{1-p_{1o}}{v};\beta_0,\tau_0,d\right)\\
&\quad+n^{-1/2}\sum_{j=1}^{n}\frac{1}{v(1-p_{1o})}f_{P|D}\left(1-\frac{1-p_{1o}}{v}|0\right)(I\{D_j=1\}-p_{1o})+o_P(1)\\
&=n^{-1/2}\sum_{j=1}^{n}V_{j,V}(v;\theta_0,\tau_0)+o_P(1).
\end{aligned}
$$

Step 3 Steps 1 and 2 imply: uniformly in $P\in\mathcal{P}$,

$$
\begin{aligned}
&\{\sqrt{n}(\hat{G}_{W|D}(w|1)-G_{W|D}(w|1),\ \hat{G}_{V|D}(v|0)-G_{V|D}(v|0))':(w,v)\in\mathcal{W}\times\mathcal{V}\}\\
&\Rightarrow\{\nu_W(w|1),\nu_V(v|1):(w,v)\in\mathcal{W}\times\mathcal{V}\},
\end{aligned}
$$

where $\{\nu_W(w|1),\nu_V(v|1):(w,v)\in\mathcal{W}\times\mathcal{V}\}$ is a vector-valued Gaussian process on $\mathcal{W}\times\mathcal{V}$ with zero mean and a covariance kernel given by $C((w_1,v_1),(w_2,v_2))$. Finally, we obtain: uniformly in $P\in\mathcal{P}$,

$$
\begin{aligned}
&\sqrt{n}(c_1\hat{\mu}_1^L+c_2\hat{\mu}_1^U+c_3\hat{\mu}_0^L+c_4\hat{\mu}_0^U-[c_1\mu_1^L+c_2\mu_1^U+c_3\mu_0^L+c_4\mu_0^U])\\
&\Rightarrow c_1\int_0^{1-P_{01}}\frac{\nu_W}{g_{W|D}}\circ G_{W|D}^{-1}(u|1)du+c_2\int_{P_{01}}^{1}\frac{\nu_W}{g_{W|D}}\circ G_{W|D}^{-1}(u|1)du\\
&\quad+c_3\int_0^{1-P_{00}}\frac{\nu_V}{g_{V|D}}\circ G_{V|D}^{-1}(u|0)du+c_4\int_{P_{00}}^{1}\frac{\nu_V}{g_{V|D}}\circ G_{V|D}^{-1}(u|0)du.
\end{aligned}
$$

Q.E.D.

Proof of Theorem 6.2: We need to show that uniformly in $P\in\mathcal{P}$,

$$
\left[c_1\int_0^{1-P_{00}}\hat{G}_{V/W|D}^{-1}(u|0)du+c_2\int_{P_{00}}^{1}\hat{G}_{V/W|D}^{-1}(u|0)du\right]
$$

is asymptotically normal for all constants c_1, c_2. It is sufficient to show that the process

$$\left\{\sqrt{n}(\hat{G}_{V/W|D}(a|0)-G_{V/W|D}(a|0)):a\in\mathcal{A}\right\}$$

converges weakly to a Gaussian process uniformly in $P\in\mathcal{P}$.

Step 1 We show: uniformly in $a\in\mathcal{A}$ and $P\in\mathcal{P}$

$$\sqrt{n}(\hat{G}_{V/W|D}(a|0)-G_{V/W|D}(a|0))=n^{-1/2}\sum_{j=1}^{n}V_{j,V/W}(a;\theta_0,\tau_0)+o_P(1).$$

Let

$$X_i(\theta,\tau)=\frac{(1-p_1)p(Z_i;\beta,\tau)}{p_1[1-p(Z_i;\beta,\tau)]}.$$

Then

$$I\left\{p(Z_i;\hat{\beta},\hat{\tau})\le\frac{a\hat{p}_1}{1-(1-a)\hat{p}_1}\right\}=I\{X_i(\hat{\theta},\hat{\tau})\le a\}.$$

Note that

$$\begin{aligned}&E_P[I\{X_i(\theta,\tau)\le x\}|D_i=0]\\&=E_P\left[I\left\{p(Z_i;\beta,\tau)\le\frac{p_1x}{1-(1-x)p_1}\right\}|D_i=0\right]\\&=F_{P|D}\left(\frac{p_1x}{1-(1-x)p_1};\beta,\tau|0\right).\end{aligned}$$

We have:

$$\begin{aligned}&\Gamma_{F,P}(x|d)[\theta-\theta_0,\tau-\tau_0]\\&=\frac{x}{[1-(1-x)p_{1o}]^2}f_{P|D}\left(\frac{p_{1o}x}{1-(1-x)p_{1o}}|0\right)(p_1-p_{1o})\\&+\Gamma_P\left(\frac{p_{1o}x}{1-(1-x)p_{1o}}|0\right)[\beta-\beta_0,\tau-\tau_0].\end{aligned}$$

Thus,

$$\begin{aligned}&\sqrt{n}(\hat{G}_{V/W|D}(a|0)-G_{V/W|D}(a|0))\\&=\sqrt{n}\left(\hat{F}_{P|D}\left(\frac{a\hat{p}_1}{1-(1-a)\hat{p}_1}|0\right)-G_{V/W|D}(a|0)\right)\\&=n^{-1/2}\sum_{j=1}^{n}\frac{1}{(1-p_{1o})}\left[I\left\{p(Z_j;\beta_0,\tau_0)\le\frac{ap_{1o}}{1-(1-a)p_{1o}}\right\}-G_{V?W|D}(v|0)\right]I\{D_j=0\}\\&+n^{-1/2}\sum_{j=1}^{n}\frac{1}{(1-p_{1o})}\psi_j\left(\frac{p_{1o}a}{1-(1-a)p_{1o}};\beta_0,\tau_0,0\right)\\&+n^{-1/2}\sum_{j=1}^{n}\frac{a}{(1-p_{1o})[1-(1-a)p_{1o}]^2}f_{P|D}\left(\frac{p_{1o}a}{1-(1-a)p_{1o}}|0\right)(I\{D_j=1\}-p_{1o})\\&+o_P(1)\\&=n^{-1/2}\sum_{j=1}^{n}V_{j,V/W}(a;\theta_0,\tau_0)+o_P(1).\end{aligned}$$

Step 2 Step 1 implies:

$$\left\{\sqrt{n}(\hat{G}_{V/W|D}(a|0)-G_{V/W|D}(a|0)):a\in\mathcal{A}\right\}$$

weakly converges to a Gaussian process $\nu_{V/W}(\cdot)$ with zero mean and covariance kernel:

$$E(V_{j,V/W}(a_1;\theta_0,\tau_0)V_{j,V/W}(a_2;\theta_0,\tau_0))$$

By the Functional Delta method, we obtain:

$$\begin{aligned}&\sqrt{n}([c_1\hat{\mu}^L_{0|1}+c_2\hat{\mu}^U_{0|1}]-[c_1\mu^L_{0|1}+c_2\mu^U_{0|1}])\\&\Rightarrow c_1\int_0^{1-P_{00}}\frac{\nu_{V/W}}{g_{V/W|D}}\circ G^{-1}_{V/W|D}(u|0)du+c_2\int_{P_{00}}^1\frac{\nu_V/W}{g_{V/W|D}}\circ G^{-1}_{V/W|D}(u|0)du.\end{aligned}$$

Q.E.D.

References

Andrews, D. W. K. 1994. "Empirical Process Methods in Econometrics." In *Handbook of Econometrics*, vol. IV, edited by R. F. Engle and D. L. McFadden. North-Holland, Amsterdam.

Andrews, D. W. K. and G. Soares. 2010. "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection." *Econometrica* 78: 119–157.

Bhattacharya, D. 2007. "Inference on Inequality from Household Survey Data." *Journal of Econometrics* 137: 674–707.

Chen, X., O. Linton, and I. van Keilegom. 2003. "Estimation of Semiparametric Models when the Criterion Function is Not Smooth." *Econometrica* 71: 1591–1608.

Chernozhukov, V., H. Hong, and E. Tamer. 2007. "Parameter Set Inference in a Class of Econometric Models." *Econometrica* 75: 1243–1284.

Cho, W. and C. F. Manski. 2008. "Cross Level/Ecological Inference." In *Oxford Handbook of Political Methodology*, edited by H. Brady, D. Collier, and J. Box-Steffensmeier, pp. 547–569. Oxford: Oxford University Press.

Cross, P. J., and C. F. Manski. 1999. "Regressions, Short and Long." Manuscript.

Cross, P. J., and C. F. Manski. 2002. "Regressions, Short and Long." *Econometrica* 70 (1): 357–368.

DiNardo, J., N. Fortin, and T. Lemieux. 1996. "Labor Market Institutions and the Distribution of Wages, 1973–1992: A Semiparametric Approach." *Econometrica* 64: 1001–1044.

Dehejia, R., and S. Wahba. 1999. "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94: 1053–1062.

Duncan, O., and B. Davis. 1953. "An Alternative to Ecological Correlation." *American Sociological Review* 18: 665–666.

Fan, Y., and S. Park. 2009. "Partial Identification of the Distribution of Treatment Effects and its Confidence Sets." *Advances in Econometrics: Nonparametric Econometric Methods* 25: 3–70.

Fan, Y., and S. Park. 2010. "Sharp Bounds on the Distribution of Treatment Effects and Their Statistical Inference." *Econometric Theory* 26: 931–951.

Fan, Y., and K. Song. 2011. "Confidence Sets for the Distribution of Treatment Effects with Covariates." Working Paper.

Fan, Y., R. Sherman, and M. Shum. 2014. "Identifying Treatment Effects under Data Combination." *Econometrica* 82: 811–822.

Firpo, S., N. Fortin, and T. Lemieux. 2010. "Decomposition Methods in Economics." In *Handbook of Labor Economics*, edited by David Card and Orley Ashenfelter, 4. New York: North-Holland.

Frank, M., R. Nelson, and B. Schweizer. 1987. "Best-Possible Bounds on the Distribution of a Sum – a Problem of Kolmogorov." *Probability Theory and Related Fields* 74: 199–211.

Goldie, C. 1977. "Convergence Theorems for Empirical Lorenz Curves and their Inverses." *Journal of Applied Probability* 9: 765–791.

Goodman, L. 1953. "Ecological Regressions and Behavior of Individuals." *American Sociological Review* 18: 663–664.

Hahn, J. 1998. "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects." *Econometrica* 66: 315–331.

Heckman, J., H. Ichimura, J. Smith, and P. Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66: 1017–1098.

Hirano, K., G. W. Imbens, and G. Ridder. 2000. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *NBER Technical Working Papers* 0251, National Bureau of Economic Research, Inc.

King, G. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton: Princeton University Press.

Linton, O., E. Maasoumi, and Y.-J. Whang. 2005. "Consistent Testing for Stochastic Dominance Under General Sampling Schemes." *Review of Economic Studies* 72: 735–765.

Linton, O., K. Song, and Y.-J. Whang. 2010. "An Improved Bootstrap Tests of Stochastic Dominance." *Journal of Econometrics* 154: 186–202.

Manski, C. F. 1990. "Nonparametric Bounds on Treatment Effects." *American Economic Review* 80: 319–323.

Rosenbaum, P. R., and D. B. Rubin. 1983a. "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome." *Journal of the Royal Statistical Society, Series B* 45: 212–218.

Rosenbaum, P. R., and D. B. Rubin. 1983b. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70: 41–55.

Stoye, J. 2009. "More on Confidence Intervals for Partially Identified Parameters." *Econometrica* 77: 1299–1315.

van der Vaart, A., and J. Wellner. 1996. *Weak Convergence and Empirical Processes*. Heidelberg: Springer Verlag.

Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press.

Supplemental Material: The online version of this article (DOI: 10.1515/jem-2015-0006) offers supplementary material, available to authorized users.